Relational Information Predicts Human Behavior and Neural Responses to Complex Social Scenes

Wenshuo Qin (wqin6@jh.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, Maryland 21218, United States

Manasi Malik (mmlik16@jhu.edu)
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

Understanding social scenes depends on tracking relational visual information, which is prioritized behaviorally and represented in the superior temporal sulcus (STS). However, computational models often overlook these cues. Here, we evaluate two social interaction recognition models, SocialGNN and RNN Edge, that explicitly incorporate relational signals—gaze direction or physical contact—and compare their predictions to human behavioral and neural responses. SocialGNN organizes video frames into a graph structure with nodes representing faces and objects, and edges encoding relational signals. RNN Edge is simpler, processing only relational information over time without node features. We found both models strongly predicted human behavioral ratings of social interactions and were comparable to state-of-the-art AI models with far less training data and simpler architectures. Both models also better predict STS responses of people watching social interaction videos than a matched visual model trained without relational cues. These findings underscore the value of integrating relational cues into computational models of social vision.

Keywords: Artificial Intelligence; Cognitive Neuroscience; Social Cognition; Neural Networks; fMRI

Introduction

Humans effortlessly recognize interactions between other agents, a skill suggested to be supported by the superior temporal sulcus (STS) (Deen, Koldewyn, Kanwisher, & Saxe, 2015; Deen, Saxe, & Kanwisher, 2020; Isik, Koldewyn, Beeler, & Kanwisher, 2017; McMahon, Bonner, & Isik, 2023) and driven by bottom-up relational cues such as gaze, proximity, and touch (Hafri & Firestone, 2021; McMahon et al., 2023; Papeo, 2020). Yet even the best computer vision systems largely ignore these cues and lag behind human social scene understanding (Garcia, McMahon, Conwell, Bonner, & Isik, 2024; Shu et al., 2021). Bridging this gap is crucial for both cognitive science and human-aligned AI. We, therefore, compared SocialGNN (Malik & Isik, 2023), a graph neural network with relational structure, with a simpler new counterpart: RNN Edge. SocialGNN feeds frame-level graphs with both node (face and object) and edge (pairwise relation like gaze and touch) features into a long short-term memory (LSTM) to capture spatial and temporal structure. RNN Edge drops the node features entirely, testing the performance of relational cues alone.

We conducted three experiments. First, we replicated and extended the results from the dataset in Malik and Isik (2023) and found that in both natural videos and animated shape stimuli, RNN Edge can predict human judgments as well as SocialGNN, indicating that representations learned over edges alone can drive accurate classification. Next, we compared both models to the human judgment of short video clips of people interactions and found they predict human judgments as well as state-of-the-art (SOTA) vision models. Finally, in fMRI data of people watching these clips, both relational models especially better predict responses in the superior temporal sulcus (STS) region compared to other brain areas. For a more detailed account of these experiments see Qin, Malik, and Isik (in press).

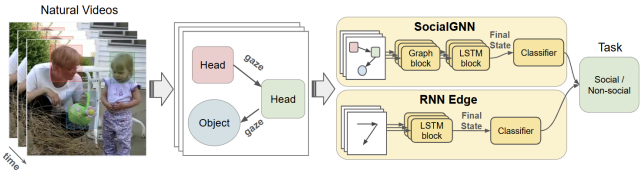

Figure 1: In natural videos, SocialGNN represents each frame as a graph with nodes (DNN embeddings of faces or objects) and directed edges (gaze), then feeds these into an LSTM and a classifier. RNN Edge uses only frame-by-frame gaze information, feeding this into an LSTM and classifier.

Methods

We extracted node and edge features for two datasets used in Malik and Isik (2023)—VACATION and PHASE—and then trained three models on these representations. VACATION is a real-world social-video collection annotated for faces, objects, and directed gaze (Fan, Wang, Huang, Tang, & Zhu, 2019); each frame yields (i) VGG-19 node embeddings reduced to 90 principal components ($\approx$ 75 % variance) and (ii) a 20-dimensional binary vector indicating who looks at whom. PHASE comprises 2-D physics-based animations in which two agents interact in friendly, neutral, or adversarial ways (Netanyahu, Shu, Katz, Barbu, & Tenenbaum, 2021); nodes encode velocity, position, size, and type, while edges are 12-dimensional binary contact vectors. We kept the original partitions: VACATION was evaluated on 20 bootstrapped splits ($\approx$ 740/215 train/test clips), and PHASE on a single 400/100 split with their respective binary and ternary labels.

Using these features, we retrained (i) SocialGNN (Malik & Isik, 2023), which ingests both node and edge features, processes each frame through a GNN, and integrates temporal context via an LSTM (L2 = 0.2) (Fig. 1); (ii) RNN Edge, a new variant that receives only the relational binary vectors, isolating edge information, thus allowing us to study the contribution of relational information (Fig. 1); and (iii) RNN Node (VisualRNN in Malik and Isik (2023)), which processes node embeddings alone, with no information from the edges.

To test generalization and whether the models capture fine-grained human ratings and fMRI responses, we evaluated the models on a separate dataset of social video responses. We used the 250 three-second two-people natural interaction clips from McMahon et al. (2023), which provides (i) online behavioral ratings along dimensions—spatial expanse (small versus large scenes), inter-agent distance, the extent to which agents are facing, the presence of object-directed actions, joint physical interactions between agents and communicative interactions; and (ii) fMRI responses from four participants covering early visual cortex (EVC), middle temporal

area (MT), posterior and anterior superior temporal sulcus (pSTS, aSTS), fusiform face area (FFA), and parahippocampal place area (PPA). Previous benchmarking has shown that SOTA models struggle to match human behavior and neural responses to these videos (Garcia et al., 2024).

Prior to model fitting, we refined the 250-clip natural-video set of McMahon et al. (2023) by (i) adding head/object bounding boxes, (ii) labeling gaze directions, and (iii) discarding 2 clips with fewer than two visible heads. Each three-second video was segmented whenever the number of visible agents changed, yielding 230 train and 51 test sub-clips; representations from sub-clips of the same parent video were averaged before regression. Model, behavioral, and fMRI features were z-scored (fit on train only) and aligned with leave-one-out ridge regression, searching seven alphas from 10^{-2} to 10^{5}. All results are benchmarked against DeiT3-L, the best vision model overall on the behavioral encoding task on the same dataset reported by Garcia et al. (2024).

Results

In our replication analysis, RNN Edge surpassed SocialGNN on VACATION, and both relational models outperformed the non-relational RNN Node model (two-tailed paired-sample permutation test, $n = 10{,}000$ resamples, $p < 0.001$ for both cases), showing that models trained with a simplified relational focus can outperform those with a non-relational focus. On PHASE, both edge-aware models, SocialGNN and RNN Edge, scored within the range of human agreement ($\approx$ 84 %) and matched the inverse-planning SIMPLE baseline, despite using orders-of-magnitude fewer parameters (data not shown for space).

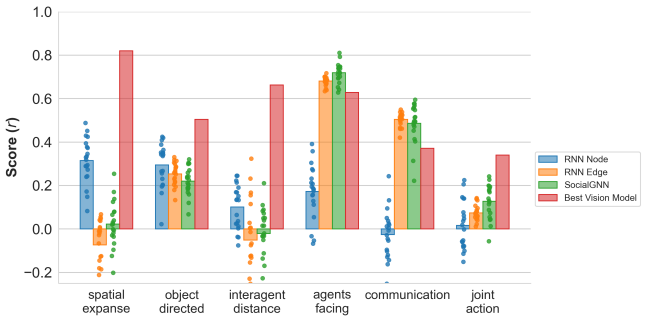

Figure 2: Behavioral Encoding. For RNN Node, RNN Edge, and SocialGNN, each dot represents a model trained using the different bootstrapped train-test splits, with 20 bootstraps per model type, and the bars denote the average performance. For the best vision model (Garcia et al., 2024), DeiT3-L, the bar is the encoding score from this single model.

In the behavioral encoding task (Fig. 2), both SocialGNN and RNN Edge achieved high scores on "agent facing" and "communicating", and permutation tests (two-tailed paired-sample, $n = 10{,}000$ resamples) showed that each edge model captured significantly more information than the control RNN Node model ($p = 0.0002$ for both dimensions). Both SocialGNN and RNN Edge also outperformed the SOTA vision-transformer baseline. By contrast, RNN Node better predicted spatial and object-centric attributes, outperforming both relational models on "spatial expanse" ($p = 0.0002$) and "inter-agent distance" ($p \leq 0.014$), and outperforming SocialGNN on "object directed" ($p \leq 0.0430$). These results highlight the strength of edge representations for social cues and node representations for spatial/object properties.

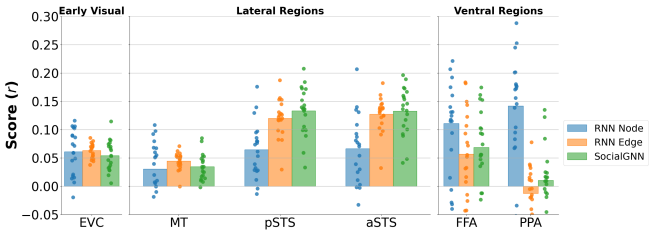

Figure 3: Neural Encoding. Each dot represents a model trained using different bootstrapped train-test splits, with 20 bootstraps per model type. The bar for each model denotes the average performance.

In the neural encoding analysis (Fig. 3), permutation tests (two-tailed paired-sample, $n = 10{,}000$) showed that SocialGNN and RNN Edge reliably outperformed RNN Node in the pSTS and aSTS (all comparisons $p < 0.01$), and the two edge-based models did not differ from each other (aSTS:$p = 0.6195$; pSTS:$p = 0.2436$), showing relational models better encode social interaction STS regions. In the ventral stream, RNN Node outperformed both edge models in the PPA ($p < 0.001$) and exhibited a small, non-significant advantage in the FFA, showing nonrelational model holds advantages in the ventral region. No reliable model differences were observed in the EVC.

Discussion

Models that focus on relational cues align more closely with human social judgments of agents facing and communication than models that ignore edge information. An extremely simple model RNN Edge matched or exceeded both the SocialGNN and much larger and more expensive SOTA vision transformers, showing that simple edge representations can explain complex social decisions. Neurally, the edge models outperformed node models in STS—providing novel evidence that this region represents relational information about social scenes—but offered no benefit in early visual or ventral areas. Overall, these results underscore the power of relational representations and highlight a flexible inductive bias to learn such representations in simple neural networks. RNN Edge performance matches that of SocialGNN, suggesting that SocialGNN may rely primarily on edge information, perhaps due to node complexity or suboptimal node features. Future works will study the edge and node representations in these models to refine their complementary roles.

Acknowledgment

We thank Emaile McMahon for facilitating access to the encoding dataset and Kathy Garcia for her help with the benchmarked DNN models. We are grateful to the members of the Isik and Bonner laboratories for helpful discussions and feedback. Computing was supported by the Advanced Research Computing at Hopkins (ARCH). This work was funded by the NIH R01MH132826 Grant awarded to Dr.Leyla Isik.

References

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015, November). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cereb Cortex*, *25*(11), 4596–4609. doi: 10.1093/cercor/bhv111

Deen, B., Saxe, R., & Kanwisher, N. (2020, November). Processing communicative facial and vocal cues in the superior temporal sulcus. *NeuroImage*, *221*, 117191. doi: 10.1016/j.neuroimage.2020.117191

Fan, L., Wang, W., Huang, S., Tang, X., & Zhu, S.-C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5724–5733).

Garcia, K., McMahon, E., Conwell, C., Bonner, M. F., & Isik, L. (2024, June). *Modeling dynamic social vision highlights gaps between deep learning and humans.* OSF. doi: 10.31234/osf.io/4mpd9

Hafri, A., & Firestone, C. (2021, June). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. doi: 10.1016/j.tics.2021.01.006

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017, October). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1714471114

Malik, M., & Isik, L. (2023, November). Relational visual representations underlie human social interaction recognition. *Nat Commun*, *14*(1), 7317. (Publisher: Nature Publishing Group) doi: 10.1038/s41467-023-43156-8

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

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021, May). PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), 845–853. (Number: 1) doi: 10.1609/aaai.v35i1.16167

Papeo, L. (2020). Twos in human visual perception. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *132*, 473–478. (Place: France Publisher: Elsevier Masson SAS) doi: 10.1016/j.cortex.2020.06.005

Qin, W., Malik, M., & Isik, L. (in press). Relational information predicts human behavior and neural responses to complex social scenes. In *Proceedings of the 47th annual conference of the cognitive science society.* (accepted)

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., ... Ullman, T. D. (2021, July). *AGENT: A Benchmark for Core Psychological Reasoning.* arXiv. (arXiv:2102.12321 [cs]) doi: 10.48550/arXiv.2102.12321