Social Inference From Relational Visual Information: An Investigation With Graph Neural Network Models

Manasi Malik & Leyla Isik, Johns Hopkins University

Humans easily recognize social interactions from visual input

Two competing computational models of human social interaction judgments are:

Bottom-Up Visual Models

+ Aligned with behavioural & neural evidence
- However, have been unsuccessful at modelling human judgments

Generative Inverse Planning Models

+ So far best match to human data
- However, computationally expensive and infeasible in natural settings

Our Hypothesis: Adding relational inductive biases to bottom-up visual models will allow them to replicate human social interaction judgments

Conclusion & Discussion

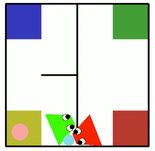
A visual model with relational inductive biases matches human social interaction judgements across both animated and natural videos.

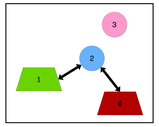
- Implication for AI systems: An added human-like inductive bias allows SocialGNN to make more human-like social judgments without incurring the computational cost of Generative inverse planning models.
- Insights into human cognition: Bottom-up visual information is sufficient for human social interaction judgements, and relational representation may underlie this ability.

METHODS

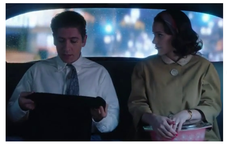
Extracting Visual Graphs from Animated and Natural Videos

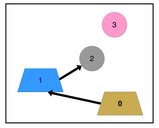
Animated Videos : PHASE dataset (Netanyahu et al., 2021)





Natural Videos : Human Gaze Communication dataset (Fan et al., 2019)





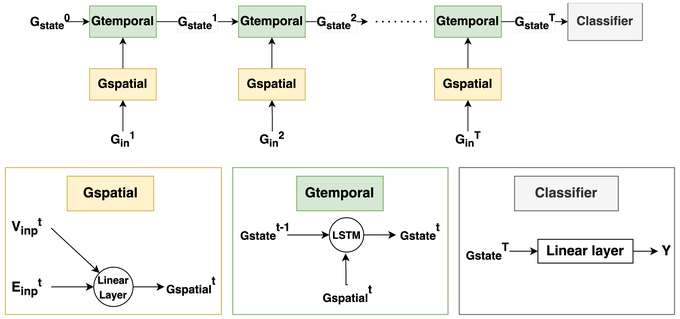
A graph (right) is created for a single frame (left) in a video for the PHASE (top) and the HGC (bottom) datasets.

Nodes: people/agents and objects in the frame

Edges : contact (PHASE) or annotated gaze direction (HGC).

All extracted Visually

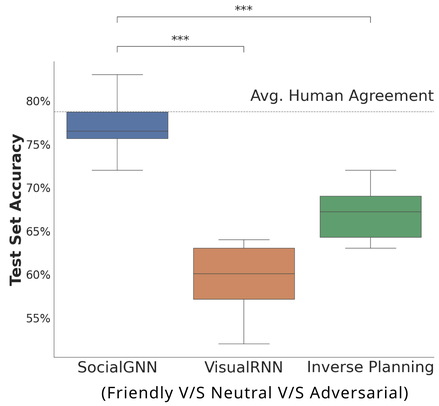
SocialGNN Architecture

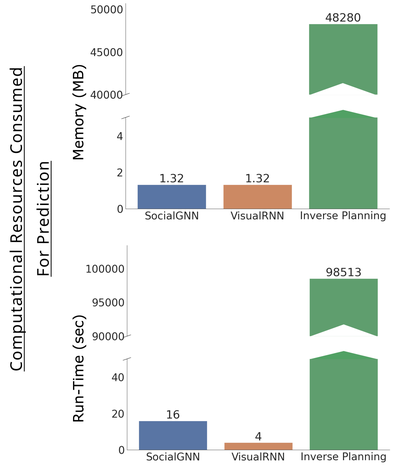


(Model developed using graph-nets library : Battaglia et al., 2018)

Control Model: VisualRNN (same RNN structure and visual input, but lacking in graph processing)

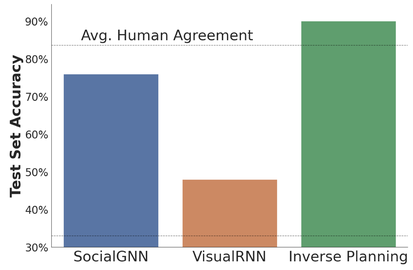
Predicting The Type Of Social Interaction In Animated Videos





Prediction On The Generalisation Set :

Videos With Novel Environment Layouts And Unseen Interactions

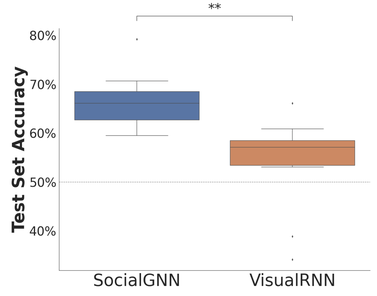


=> Relational Graphical Representations Allow Purely Visual Models To Make Human-Like Social Judgements

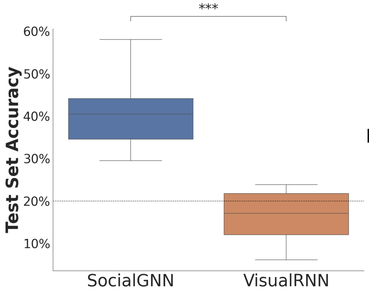
=> Much Less Computationally Expensive Than Generative Inverse Planning Models

Predicting The Type Of Social Interaction In Natural Videos

Presence Of Social Interaction



Type Of Social Interaction



=> SocialGNN Architecture Generalizes To Natural Stimuli