Evaluating Object Spatial Relationships in Human and Machine Vision

Georgina Woo, RT Pramod, Nancy Kanwisher
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Introduction

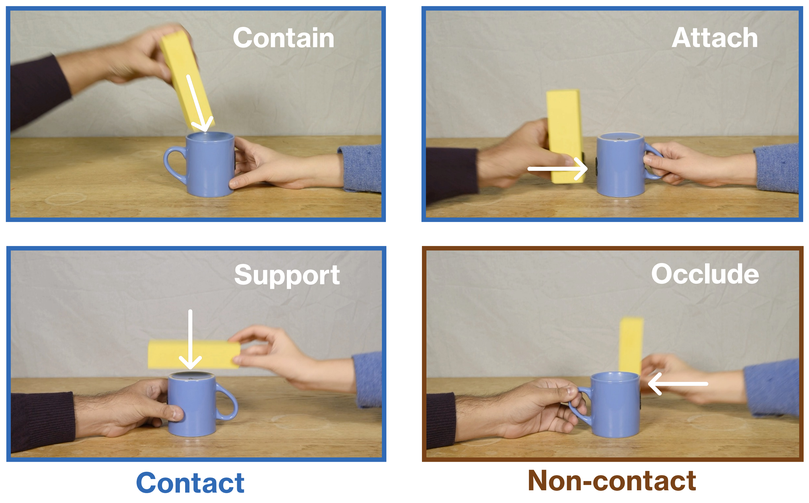
Successful interaction with the world requires **understanding the physical structure** of the scene in front of us.

This includes extracting the relationships between objects like containment, support, attachment, and occlusion. While people are quite accurate at recognizing these relationships and using them to make predictions, we still don't understand the computational mechanisms underlying such behavior.

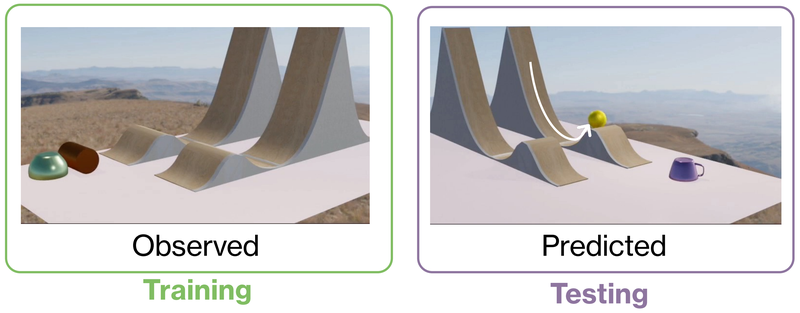
In this study, we test various **deep learning models** as computational hypotheses of **object relation understanding**.

Methods

Experiment 1 & 2: Contact Classification and Detection



Experiment 3: Contact Prediction



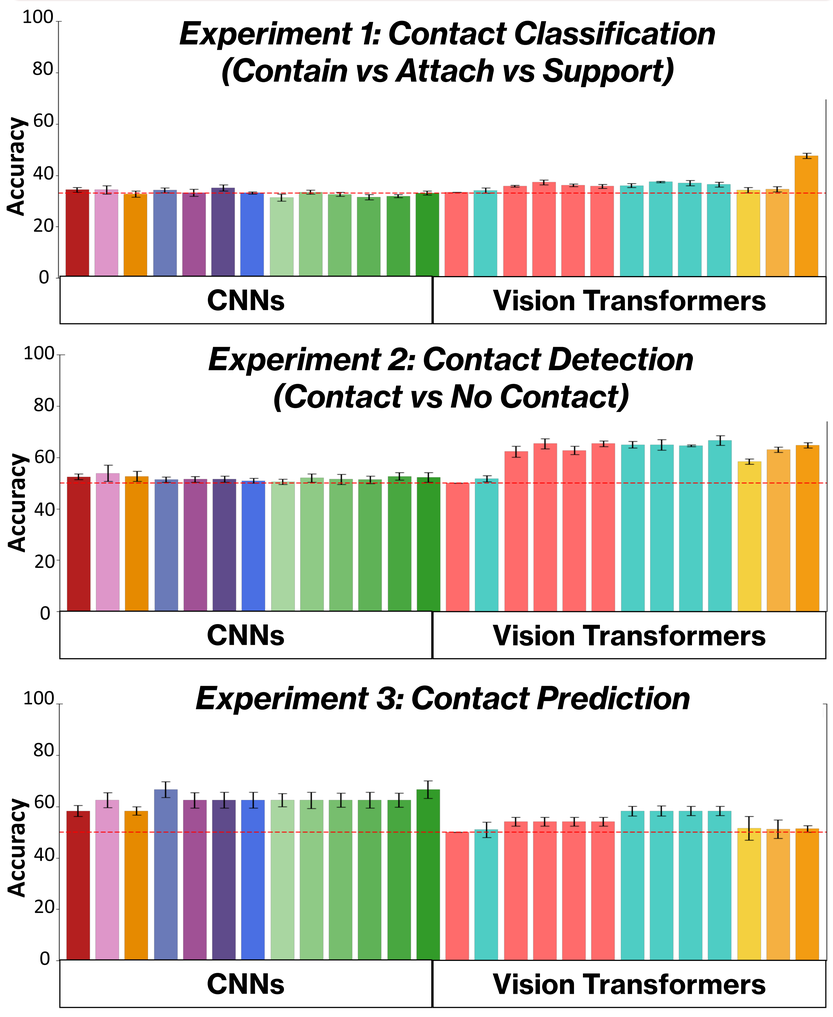
Video Trained Models

- 13 CNNs trained on *Kinetics-400*
- 11 Vision Transformers trained on *Kinetics-700, Physion,* and *VideoMix2M*

We trained linear SVM classifiers on model features for **contact classification, contact detection and future contact prediction**, and tested the classifiers on different scenarios.



Evaluating Spatial Reasoning in Videos



Evaluating Spatial Reasoning in Images

Considering poor model performance on videos, we made the task simpler by removing dynamics and testing the models on static images.

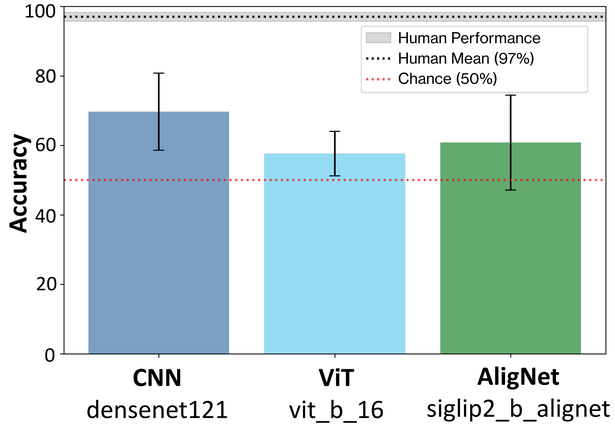
Example Trial



Human Behavioral Benchmark

Participants (N=40) completed 320 matching trials, selecting which of two alternatives shared the same spatial relationship. Humans achieved 97% accuracy across five relations (contain, support, attach, occlude, cover).

Best Model Performance by Architecture Type



Conclusion and further study

Although some computational models perform above chance on relational tasks, they fail on many scenarios obvious to humans, suggesting that standard deep learning methods alone may be **insufficient to fully capture human physical scene understanding**. Future work will examine whether fine-tuning improves model-human alignment, develop a large-scale video dataset for dynamic relational reasoning, and make these datasets publicly available to support development of better spatial reasoning models.