Introduction
Robust communicative interaction between humans and computers requires the following three capabilities:
1. Recognition and generation within multiple modalities, e.g., language, gesture, vision, action;
2. Understanding of contextual grounding and co-situatedness in conversation;
3. Appreciation of consequences of actions taken throughout the dialogue.

Central to these is “semantically grounding” a concept to a situation;
• Certain modalities are better at grounding certain types of information
  (e.g., deixis to locations, language to attributives or concept labels).

“Multimodal linking” is insufficient
Situated grounding entails knowledge of entities in context
("common ground")

“What am I pointing at?”

Validating a Situated Grounding Model
• Model uses a CNN to predict the nearest known sample to the current situation, and an LSTM to generate the most likely sequence of moves to approach it.
• As the structure approaches completion, both these predictions should get less uncertain (lower cross-entropy loss):
  • The closest target example should become clearer, as should the moves needed to get there.
• Validation: measure the training loss while increasing the size of the input to each network.
  • i.e., with 1 relation as input, remaining 19 relations should be very hard to predict; with 19 relations as input, remaining 1 relation needed should be very evident.

Grounding Novel Semantics
• Generating new instances is only part of “grounding”;
• Agents must also be able to recognize and classify learned concepts.
• We treat this as constraint satisfaction and inference.
• Approaches: weighted constraint satisfaction, POMDP, Qualitative Constraint Network
• QCN approach uses combined qualitative spatial relations with interval algebra distinctions;
  • e.g., Externally Connected (touching) vs. Disconnected
• Given structural components, what relations between components satisfy the constraints that appear in the learned samples and generated examples?

Staircase with components marked

Situatedness goes beyond visual grounding. It is a true multimodal approach to demonstrating meaning. Demonstrating knowledge ensures shared understanding.

With a semantic scaffold to transform quantitative values into qualitative values, situated simulation affords tractable language understanding.