Cooperative Vision-and-Dialog Navigation (CVDN)

A big dataset of human-human dialog! Training navigation agents! A demo interface! ↓

Visible to both Navigator and Oracle

Into the hall or the office? <NAV> Into the hall. Follow it to a living room. <NAV> Should I go upstairs? Yah, head up the stairs.

Visible Only to the Oracle

Okay, left or right from here? <NAV> Left into the bedroom. Cross it towards another door.

Human Navigator paths are longer than shortest path planner routes, resulting in much longer paths than in the comparable Room-to-Room dataset.

Human Oracles use more words than Navigators, and dialogs have on average 3-4 question-answer exchanges each.

The average percent of Dialogs, as well as individual Navigator and Oracle utterances, exhibiting each phenomena out of 100 hand-annotated dialogs.

Dataset

CVDN is the first dataset to include two-sided dialogs held in natural language, with the initial navigation instruction being both ambiguous (Amb) and underspecified (Unders), and situated in a photorealistic, visual navigation environment viewed by both speakers.

Navigation Task

For each question-answer exchange, we task an agent with navigating towards the goal given the dialog so far. We can use the path taken by the Navigator, shown to the Oracle, or a mix as supervision.

Evaluation

Average agent progress towards the goal location when trained using different path end nodes for supervision. Among ablations, bold indicates most progress by language input, and blue indicates most progress by supervision signal.

Using all dialog history significantly outperforms unimodal ablations in unseen environments. Using all dialog history, rather than just the last question or question-answer exchange, is needed to achieve statistically significantly better performance than using the target object alone in unseen test environments. Dialog history is beneficial for understanding the context of the latest navigation instruction. Models trained with mixed supervision always statistically significantly outperform those trained with oracle or navigator supervision. Using human demonstrations only when they appear trustworthy increases agent progress towards the goal.