Vision-based Navigation with Language-based Assistance via Imitation Learning with Indirect Intervention

Khanh Nguyen, Debadepta Dey, Chris Brockett, and Bill Dolan

lknguyen@cs.umd.edu  {dedey.Chris.Brockett,bilidol}@microsoft.com  Project Github: debadepta/vnla

We incorporate into a vision-based navigation agent the abilities to
- Determine when it needs additional assistance
- Query an expert advisor for language subgoals
- Interpret and execute language subgoals to make progress

Imitation learning with Indirect Intervention (I3L)
- Direct intervention: advisor takes over control from agent (e.g., physically steer a self-driving car to force it to turn)
- Indirect intervention: advisor augments environment with extra information to influence agent’s decisions (e.g., verbally order a self-driving car to slow down)

**When to request help?**

Agent is trained to imitate a help-request teacher, who decides to request help when:
1. Agent deviates too far from the optimal path:
   \[ \text{distance}(v_t, \text{Path}_{\text{optimal}}) > \delta \]
   where \( v_t \) is the agent’s location.
2. Agent has reached the goal but still wants to move:
   \[ v_t \in V_{\text{goal}} \land \arg\max(P_t^{\text{nav}}) = \langle \text{go forward} \rangle \]
   where \( V_{\text{goal}} \) is the goal set, and \( P_t^{\text{nav}} \) is the final navigation distribution.
3. Agent is uncertain:
   \[ H[P_t^{\text{nav}}] - H[\text{UniformDist}] < \epsilon \]
   where \( H \) is the entropy function, and \( P_t^{\text{nav}} \) is the tentative navigation distribution.
4. Agent stays at the current location for too long:
   \[ v_t = v_{t-1} = \cdots = v_{t-n+1} \]
   The help-request budget is greater than or equal to the number of remaining steps.

Evaluate on previously seen and unseen environments. Dataset: 95K train, 5K val/test seen, 5K val/test unseen.