Semantic Visual Navigation by Watching YouTube Videos

Matthew Chang       Arjun Gupta       Saurabh Gupta

University of Illinois at Urbana-Champaign
Finding a bathroom in a new restaurant

This paper learns such cues for finding everyday objects (bed, chair, couches, tables, toilets) in novel indoor environments.
**Current Paradigm**

Learning via Direct Interaction (Reinforcement Learning)

- High sample complexity
- Sim2Real gap

**This Paper**

Mining Spatial Co-occurrences in Real Estate Tours from YouTube

- Passive data already available on Internet
  + Visually Diverse
Challenges in Using Such Videos

- Videos don’t come with action labels
  \[\implies\] Action Grounding via an Inverse Model [1]
- Goals and intents are not known
  \[\implies\] Use off-the-shelf Object Detectors to label frames with desired objects
- Depicted trajectories may not be optimal
  \[\implies\] Use Q-learning to learn optimal behavior from sub-optimal data [2]

Value Learning from Videos

a) Action Grounding

- Inverse Model
  - built by executing random actions on robot

b) Goal Labeling

- Object Detector
  - trained on COCO

- Value function that uses implicitly learns semantic cues for seeking objects in novel indoor environments

Q-Learning Quadruple

$f(I, c) = \max_a Q^*(I, a, c)$

- Real Estate Tour from YouTube

- c) Q-Learning

- $I_t$ and $I_{t+1}$
Learned Value Function

\[ f(I, c) \approx \text{nearness to goal} \]

Value function predicts a proxy for nearness to a goal object for a given image.
Learned Value Function

\[ f(I, c) \approx \text{nearness to goal} \]

Value function predicts a proxy for nearness to a goal object for a given image.
Using Learned Values for Semantic Navigation

Hierarchical Policy

High-Level Policy
- Decides where to go next and emits short-term goal
- Builds a topological map [1] that stores values predicted by $f(I, c)$ at different locations in different directions
- Samples most promising direction, and passes $\Delta$Pose to Low-Level Policy

Low-Level Policy
- Executes actions to achieve short-term goal
- Incrementally builds occupancy map from depth camera
- Uses Fast-Marching Method for path planning to get actions to execute
- Return control on success or failure

Experiments

• Real estate videos from a newly collected *YouTube House Tours Dataset*
  • 1387 Videos, 119 Hours, 550K transition tuples
• Simulated Robot in Visually Realistic Simulation Environment (Gibson Environments in AI Habitat)
• **Action Space**: Forward (0.25m), Turn Left (30°), Turn Right (30°)
• **Task**: Find object of interest (bed, chair, couches, tables, toilets) in novel indoor environments.
• **Metrics**:
  • SPL (measures path efficiency, higher is better)
  • Success (higher is better)
YouTube House Tours Dataset
YouTube House Tours Dataset
(1387 videos, 119 Hours)
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Supervision</th>
<th>Oracle Stop</th>
<th>Policy Stop (using $D_{coco}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SPL</td>
<td>Success (SR)</td>
</tr>
<tr>
<td>Topological Exploration</td>
<td>-</td>
<td>0.30 ± 0.02</td>
<td>0.67 ± 0.02</td>
</tr>
<tr>
<td>Detection Seeker</td>
<td>-</td>
<td>0.46 ± 0.02</td>
<td>0.75 ± 0.02</td>
</tr>
<tr>
<td>RL (RGB-D ResNet+3CNN)</td>
<td>100K ($\mathcal{E}_{\text{train}}$)</td>
<td>Sparse</td>
<td>-</td>
</tr>
<tr>
<td>RL (RGB-D ResNet+3CNN)</td>
<td>1M ($\mathcal{E}<em>{\text{train}} \cup \mathcal{E}</em>{\text{video}}$)</td>
<td>Dense</td>
<td>-</td>
</tr>
<tr>
<td>RL (RGB-D 3CNN)</td>
<td>38M ($\mathcal{E}<em>{\text{train}} \cup \mathcal{E}</em>{\text{video}}$)</td>
<td>Dense</td>
<td>-</td>
</tr>
<tr>
<td>RL (RGB ResNet)</td>
<td>20M ($\mathcal{E}_{\text{train}}$)</td>
<td>Dense</td>
<td>-</td>
</tr>
<tr>
<td>RL (Depth 3CNN)</td>
<td>38M ($\mathcal{E}_{\text{train}}$)</td>
<td>Dense</td>
<td>-</td>
</tr>
<tr>
<td>Behavior Cloning</td>
<td>40K ($\mathcal{E}_{\text{train}}$)</td>
<td>-</td>
<td>$\mathcal{V}_{yt}$</td>
</tr>
<tr>
<td>Behavior Cloning + RL</td>
<td>12M ($\mathcal{E}_{\text{train}}$)</td>
<td>Dense</td>
<td>$\mathcal{V}_{yt}$</td>
</tr>
<tr>
<td>Our (Value Learning from Videos)</td>
<td>40K ($\mathcal{E}_{\text{train}}$)</td>
<td>-</td>
<td>$\mathcal{V}_{yt}$</td>
</tr>
</tbody>
</table>

- Better than strong exploration baselines
- Stronger than even RL methods trained with dense rewards with 250x more interaction samples and 6x more environments with direct interaction access
- Stronger than behavior cloning on videos and behavior cloning + RL
Summary

• Developed a technique to learn from videos

• Learned a goal seeking value function via Q-learning

• Utilized the learned value function in a hierarchal navigation policy for object goal navigation

• Code, data and models available on project webpage
  https://matthewchang.github.io/value-learning-from-videos/

Thank You