Semantic Visual Navigation by Watching YouTube Videos

Matthew Chang Arjun Gupta Saurabh Gupta

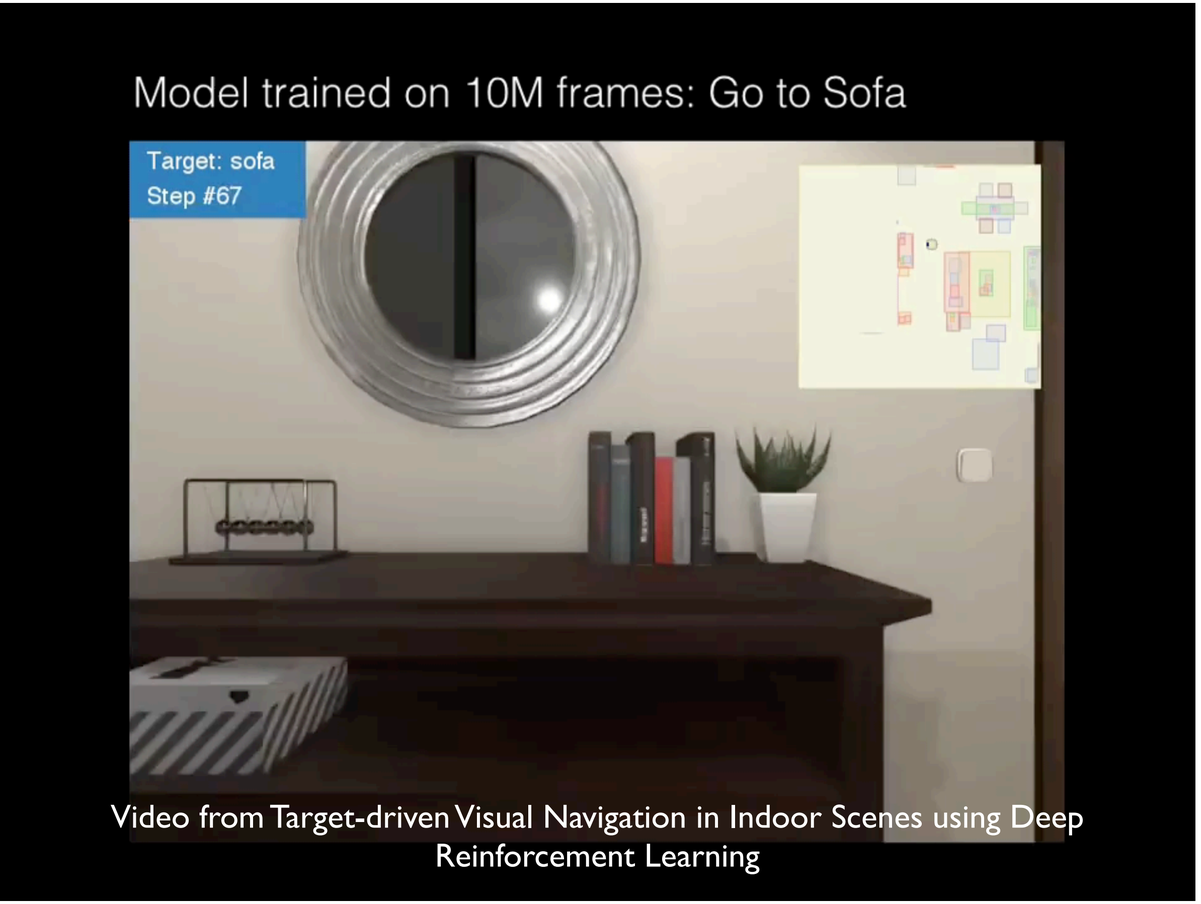
University of Illinois at Urbana-Champaign

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)



Video from Target-driven Visual Navigation in Indoor Scenes using Deep 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

  $\Longrightarrow$ Action Grounding via an Inverse Model [1]

- Goals and intents are not known

  $\Longrightarrow$ Use off-the-shelf Object Detectors to label frames with desired objects

- Depicted trajectories may not be optimal

  $\Longrightarrow$ Use Q-learning to learn optimal behavior from sub-optimal data [2]

[1] A. Kumar et al. Learning navigation subroutines by watching videos. In *CoRL*, 2019.

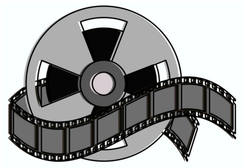
[2] Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Value Learning from Videos

a) Action Grounding

Inverse Model

built by executing random actions on robot



Real Estate Tour from YouTube

Object Detector

trained on COCO

r^c_{t+1}

b) Goal Labeling

Q-Learning Quadruple

I_t I_{t+1}

c) Q-Learning

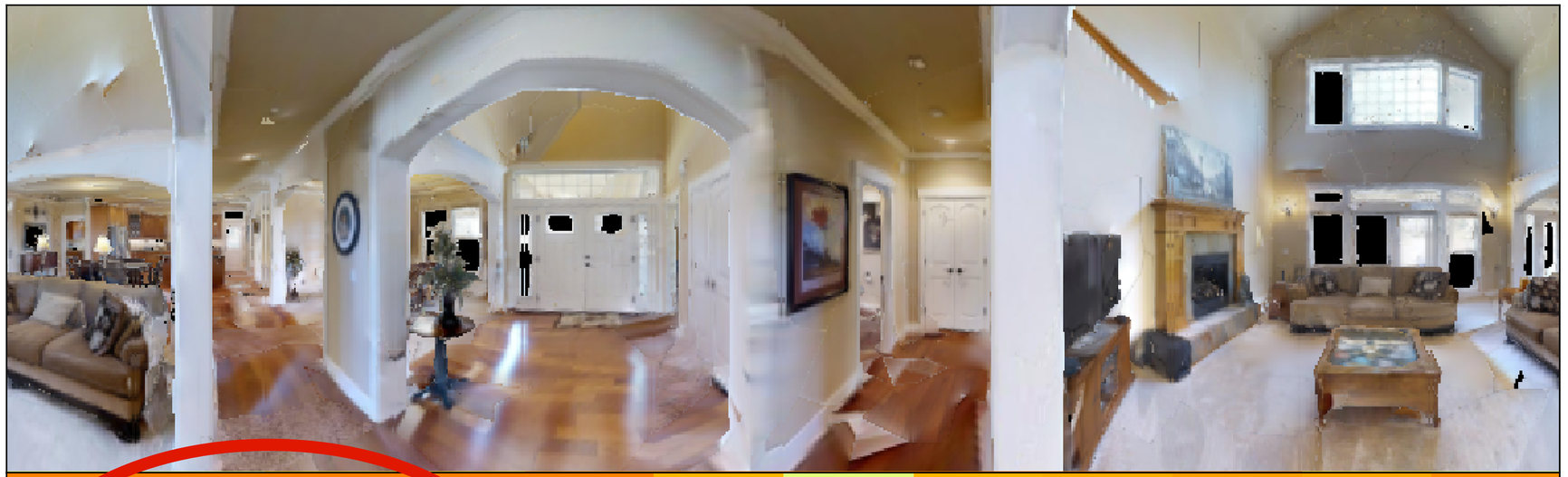
$$f(I, c) = \max_a Q^*(I, a, c)$$

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

Learned Value Function

$f(I, c) \approx$ nearness to goal

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

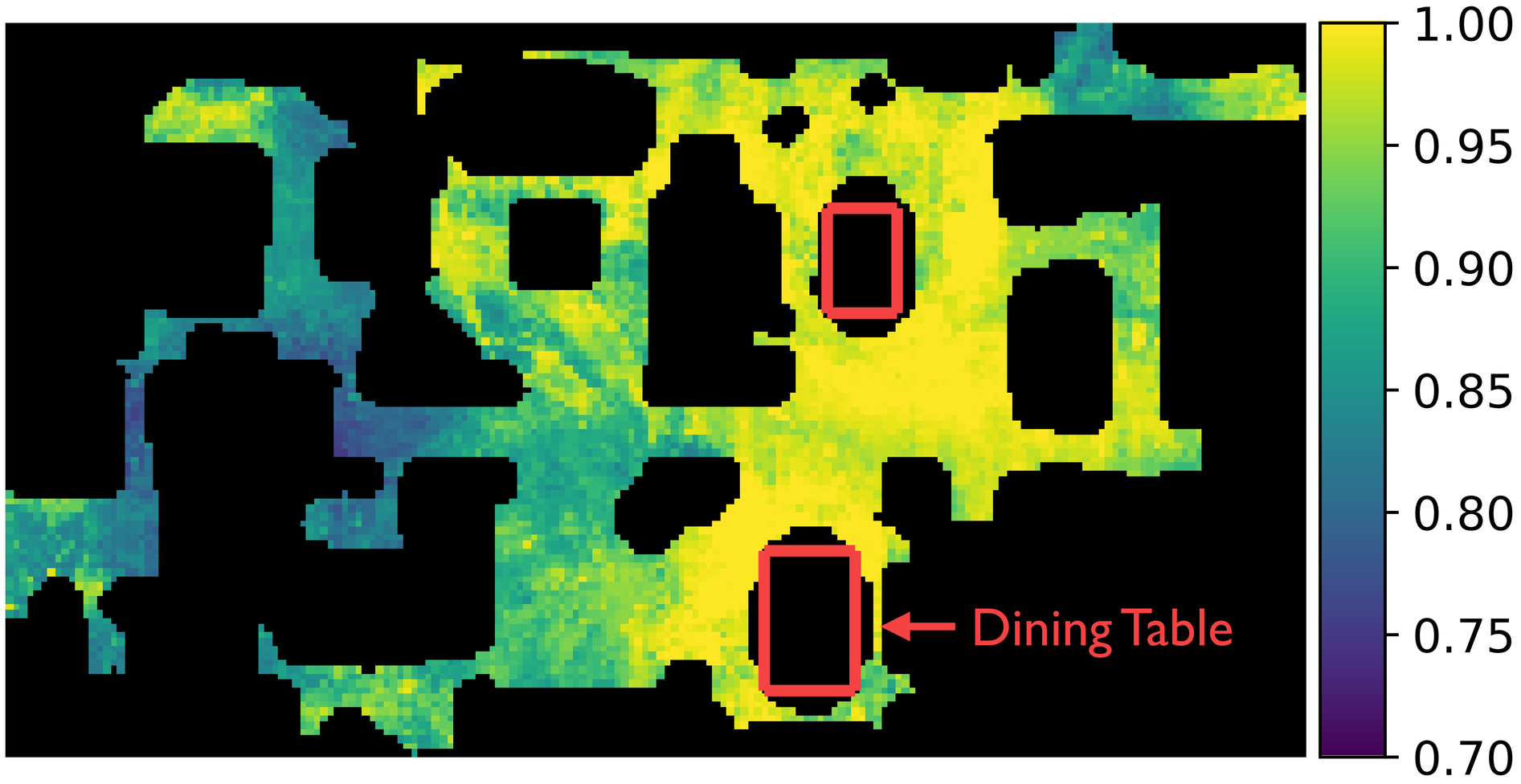


Chair	[illegible]	0.99	0.99	1.00	0.99	0.96	0.92	0.96	0.97	0.98	0.97	0.99
Couch	0.99	0.95	0.84	0.80	0.82	0.82	0.80	0.84	0.87	0.90	0.94	0.99
D. Table	0.87	0.97	0.99	1.01	0.92	0.88	0.82	[illegible]	0.85	0.85	0.84	0.83
Bed	0.78	[illegible]	[illegible]	0.80	0.83	0.83	0.84	0.84	0.84	0.83	0.80	0.78
Toilet	0.62	0.63	0.65	0.63	0.71	0.68	0.71	0.71	0.71	0.66	0.63	0.62

Learned Value Function

$f(I, c) \approx$ nearness to goal

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

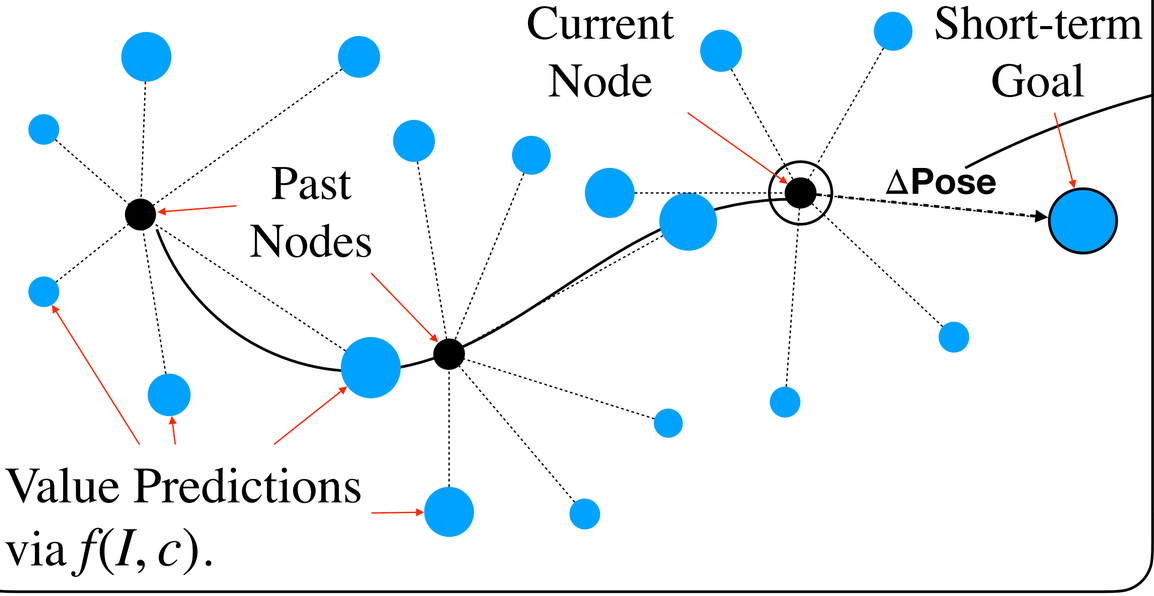


Using Learned Values for Semantic Navigation

Hierarchical Policy

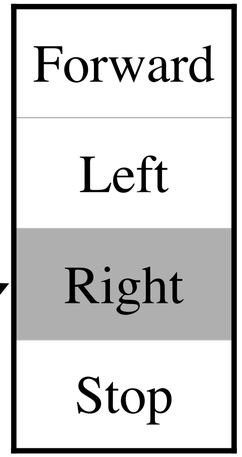
High-Level Policy

- 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 ΔPose 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

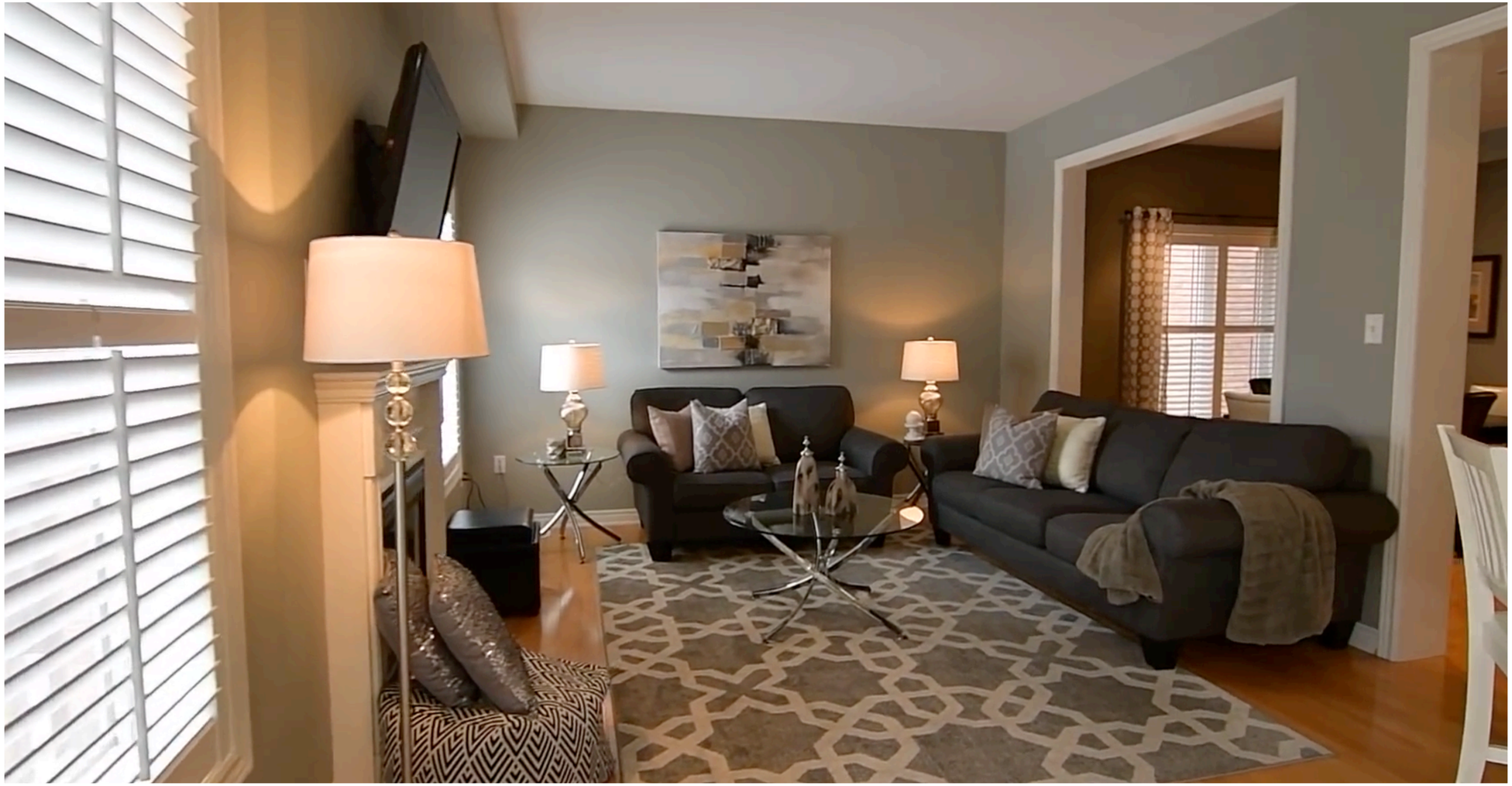


[1] D. S. Chaplot et al. Neural topological slam for visual navigation. In *CVPR*, 2020.

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

Episode automatically ends when agent reaches goal object

Agent emits STOP action to declare it has reached goal object

Method	Training Supervision			Oracle Stop		Policy Stop (using D_{coco})	
	# Active Frames	Reward	Other	SPL	Success (SR)	SPL	Success (SR)
Topological Exploration	-	-	-	0.30 ± 0.02	0.67 ± 0.02	0.13 ± 0.01	0.29 ± 0.02
Detection Seeker	-	-	-	0.46 ± 0.02	0.75 ± 0.02	0.19 ± 0.02	0.37 ± 0.02
RL (RGB-D ResNet+3CNN)	100K ($\mathcal{E}_{\text{train}}$)	Sparse	-	0.17 ± 0.01	0.37 ± 0.02		
RL (RGB-D ResNet+3CNN)	10M ($\mathcal{E}_{\text{train}} \cup \mathcal{E}_{\text{video}}$)	Dense	-	0.26 ± 0.02	0.54 ± 0.02		
RL (RGB-D 3CNN)	38M ($\mathcal{E}_{\text{train}} \cup \mathcal{E}_{\text{video}}$)	Dense	-	0.28 ± 0.02	0.57 ± 0.03		
RL (RGB ResNet)	20M ($\mathcal{E}_{\text{train}}$)	Dense	-	0.29 ± 0.02	0.56 ± 0.03	0.08 ± 0.01	0.21 ± 0.02
RL (Depth 3CNN)	38M ($\mathcal{E}_{\text{train}}$)	Dense	-	0.25 ± 0.02	0.52 ± 0.02		
Behavior Cloning	40K ($\mathcal{E}_{\text{train}}$)	-	$\hat{\mathcal{V}}_{\text{yt}}$	0.25 ± 0.02	0.53 ± 0.03	0.08 ± 0.01	0.20 ± 0.02
Behavior Cloning + RL	12M ($\mathcal{E}_{\text{train}}$)	Dense	$\hat{\mathcal{V}}_{\text{yt}}$	0.24 ± 0.02	0.58 ± 0.02		
Our (Value Learning from Videos)	40K ($\mathcal{E}_{\text{train}}$)	-	$\hat{\mathcal{V}}_{\text{yt}}$	$\mathbf{0.53} \pm 0.02$	$\mathbf{0.79} \pm 0.02$	$\mathbf{0.22} \pm 0.02$	$\mathbf{0.39} \pm 0.03$

- 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