Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction

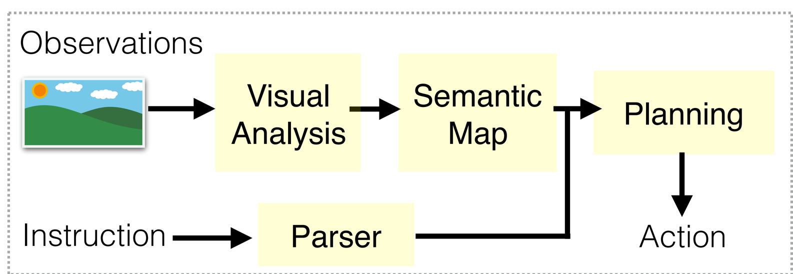
Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi

Problem and Approach

Goal: Map instructions to actions

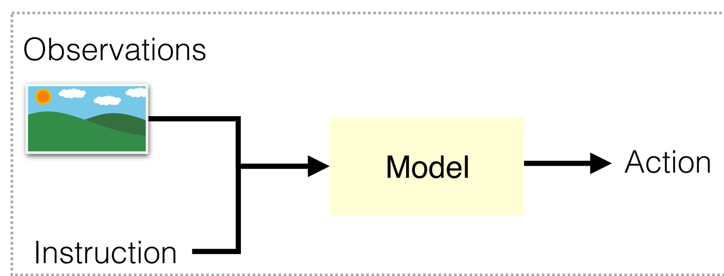
Common approaches:

Modular Engineering



Hard to scale engineering

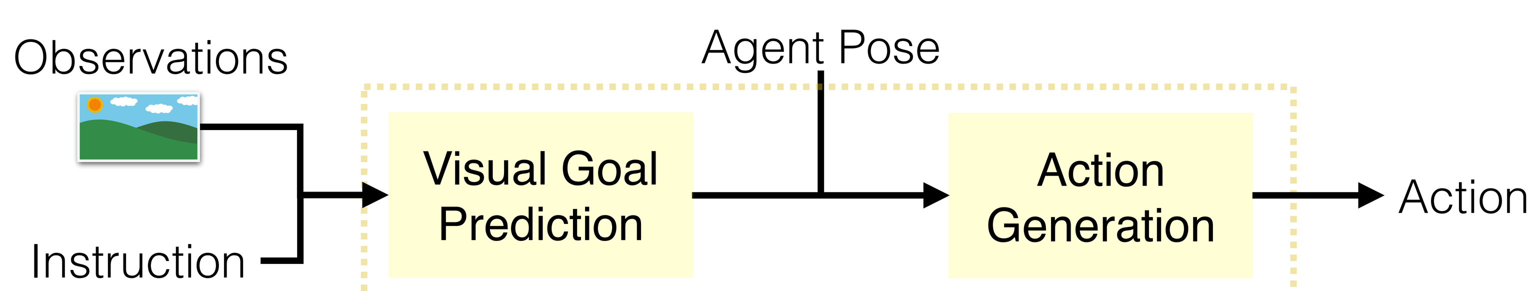
Representation Learning



Opaque blackbox reasoning

Our Approach: Visual Goal Prediction Model

A single model that decouples the problem into predicting a visual goal representation (where) and taking actions to accomplish it (how).

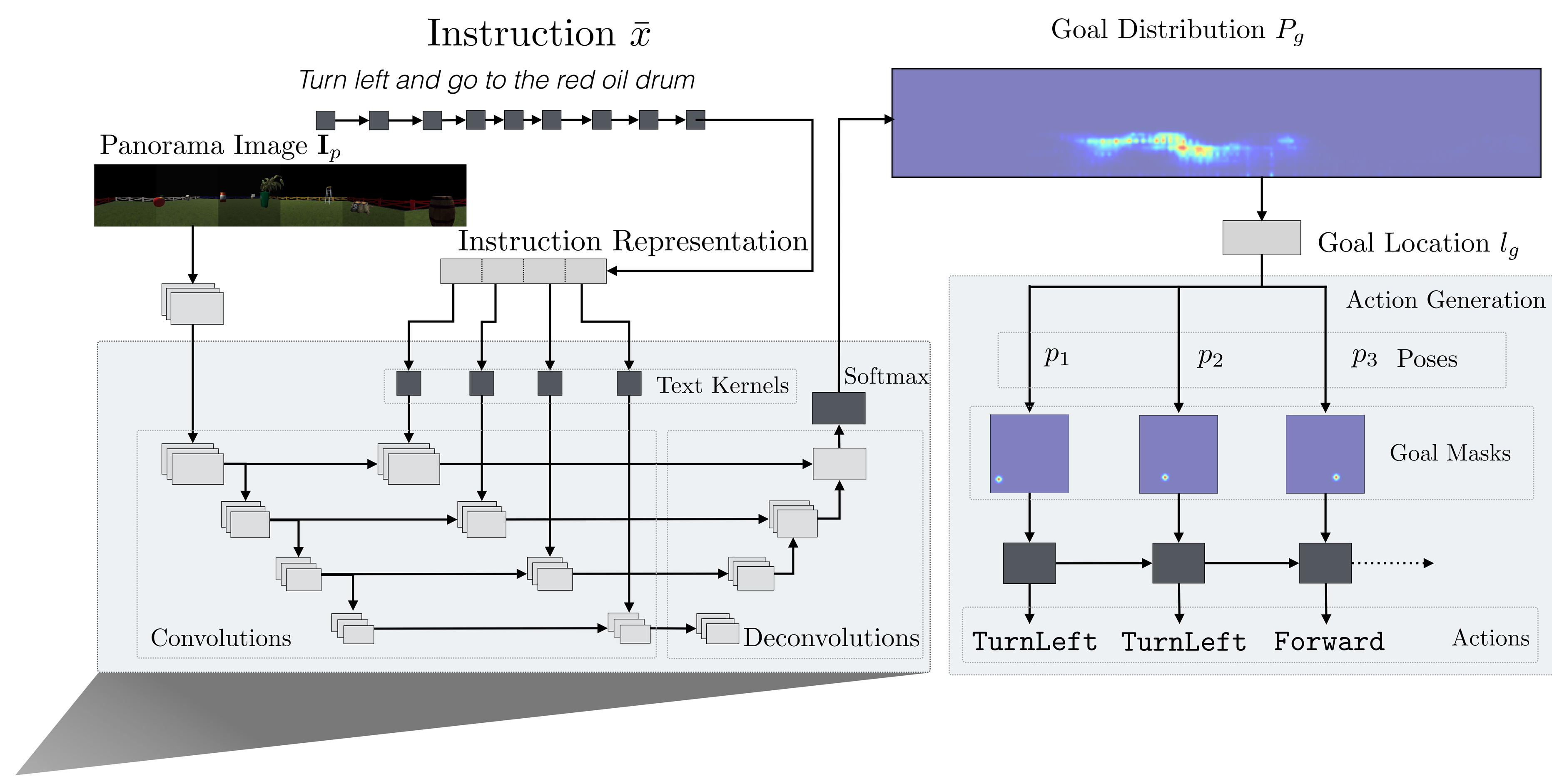


Advantages

1. **Safety and Interpretability**
   Model can predict the goal visually before taking any actions.
2. **Simplifies Learning**
   Allows training action generation in a language agnostic manner which makes learning easier.

Visual Goal Prediction Model

- Our model consists of goal prediction and action generation.
- Given a panorama of the local surrounding, we generate probability distribution over pixels representing the goal.



LingUNet:

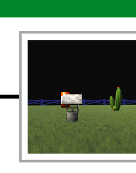
- Language-conditioned image-to-image mapping.
- Visual reasoning at multiple image scales using text-based convolutions.

Action Generation:

- Project the mode of the goal distribution to a goal location in the real world.
- Generate actions using the agent's pose and the goal location.

Two Stage Learning

- Our approach enables training the goal prediction and action generation using different learning algorithms.
- We train goal prediction using supervised learning and action generation using policy gradient in a contextual bandit setting.



Turn left and go to the red oil drum

For each example:

(1) Update the goal prediction using supervised learning.

- Given the gold goal pixel g_p maximize its log probability.

Learning rate, Parameters

$$\theta = \theta + \mu \nabla \log P_g(g_p \mid \mathbf{I}_p, \bar{x})$$

(2) Update the action generation using contextual bandit learning.

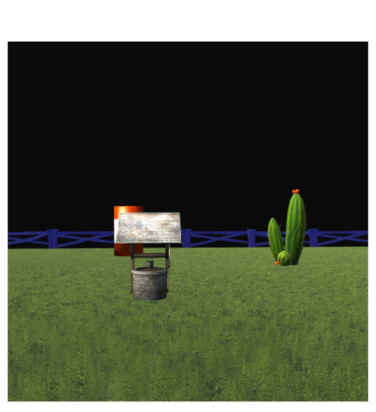
- Given the gold goal location l_g, sample actions using the policy π.
- Perform sample-efficient contextual bandit update with shaped reward (Agarwal et al., 2014, Misra et al., 2017).

Learning rate, Reward, Episode length

$$\theta = \theta + \frac{\eta}{T} \sum_{t=1}^{T} \nabla \log \pi(a_t \mid l_g, p_1, \cdots, p_t) r_t$$

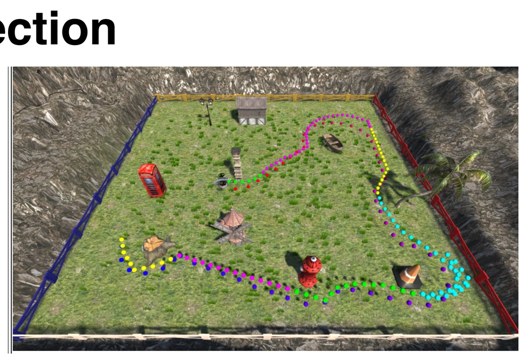
Two New Benchmarks

LANI: Navigation in an open space between landmarks. (28,204 instructions)



"After reaching the hydrant head towards the blue fence and pass towards the right side of the well."

Data Collection

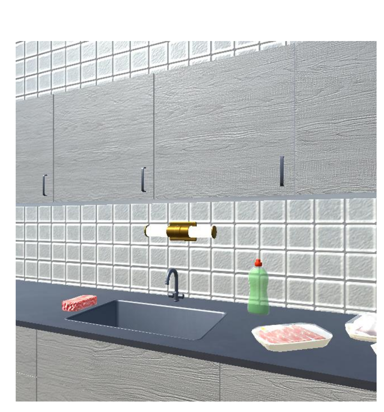


[Go around the pillar on the right hand side] [and head towards the boat, circling around it clockwise.] [When you are facing the tree, walk towards it, and the pass on the right hand side,] [and the left hand side of the cone. Circle around the cone,] [and then walk past the hydrant on your right,] [and the the tree stump.] [Circle around the stump and then stop right behind it.]

- Collect data using Amazon Mechanical Turk.
- Workers are shown a path and write instructions.
- Other workers control the agent to generate gold demonstration for every instruction.

Code, dataset and simulators available at: https://github.com/clic-lab/ciff

CHAI: Navigation and manipulation in a 3D house. (13,729 instructions)



"Put the cereal, the sponge, and the dishwashing soap into the cupboard above the sink."

- CHALET simulator (Yan et al. 2018).
- Similar data collection process to LANI.

Dataset Statistics

	LANI	CHAI
No. of paragraphs	6,000	1,596
Mean instructions per paragraph	4.7	7.70
Mean action per instruction	24.6	54.5
Mean tokens per instruction	12.1	8.4
Vocabulary Size	2,292	1,018

Corpus Analysis

	Count	
Category	**LANI**	**CHAI**
Spatial relations (locations)	123	52
Conjunctions of locations	36	5
Coordination of sub-goals	65	68
Trajectory constraints	94	0
Co-reference	32	18

200 examples manually labeled.

Results

Test Results

	LANI	CHAI	
System	**SD**	**SD**	**MA**
Stop	15.2	3.6	39.8
Misra et al. 2017	10.2	3.6	36.8
Chaplot et al. 2018	8.8	3.6	39.7
Our Approach	**8.4**	**3.3**	**40.0**

SD: Stop Distance; MA: Manipulation Accuracy

Model Ablations (DEV)

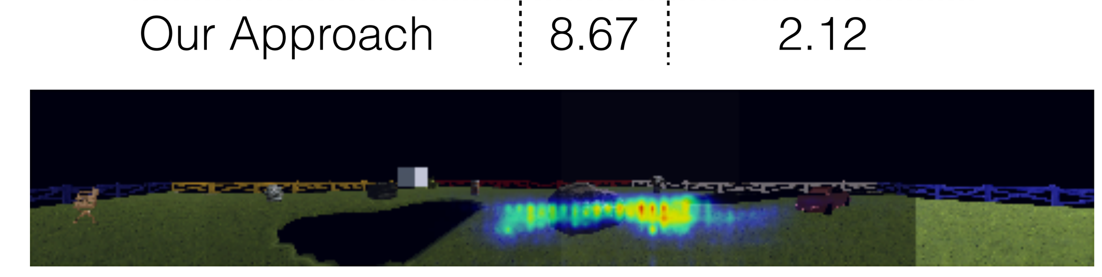
	LANI	CHAI	
Ablations	**SD**	**SD**	**MA**
Our Approach	8.65	2.75	37.53
without RNN	9.21	3.75	37.43
without language	10.65	3.22	37.53
with joint learning	11.54	2.99	36.90
with oracle goals	2.13	2.19	41.07

Linguistically-driven Analysis

- **LANI:** Temporal coordination of sub-goals and co-reference reduced the performance.
- **CHAI:** Spatial relations reduced the performance.
- Other linguistic categories did not significantly influence performance (two-sided t-test).

Visual Goal Prediction Performance

System	LANI	CHAI
Center Pixel	12.0	3.41
Janner et al. 2018	9.61	2.81
Our Approach	8.67	2.12



curve around big rock keeping it to your left .

walk over to the cabinets and open the cabinet doors up