Language Guided Meta-Control for Embodied Instruction Following

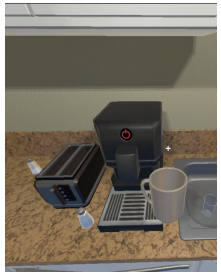
Divyam Goel[1] Kunal Pratap Singh[2] Jonghyun Choi[3]

[1] भारतीय प्रौद्योगिकी संस्थान रूड़की Indian Institute of Technology Roorkee
[2] AI2 Allen Institute for AI [3] YONSEI UNIVERSITY

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Problem Setup

Given a **human-human dialogue history** and **egocentric RGB frames**, complete tasks by inferring a **sequence of primitive actions** to achieve a goal environment state.



[High level instruction]

Can you make me a cup of coffee please.

[Dialogue History]

Where can I find a mug?

There is a mug in the top cupboard left of the fridge.

I can't seem to see a mug.

Oh sorry! I meant right of the fridge.

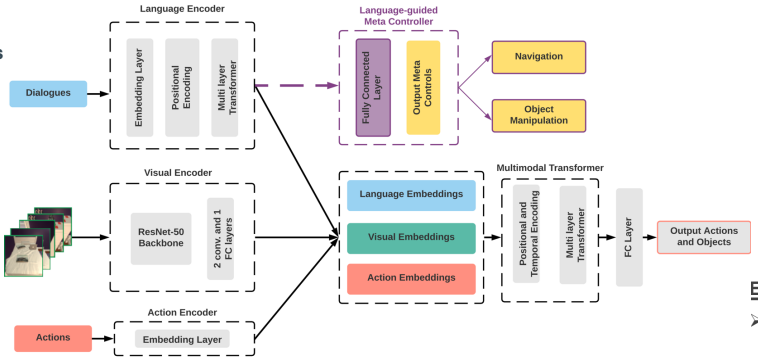
I've made a coffee.

Motivation

- Natural language instructions without additional supervision can result in unresolved ambiguities and a limited understanding of the final goal
  → **obtain clarification by learning from human-human dialogues**

- Existing frameworks are prone to underutilization of signals arising from the natural language instructions
  → **develop robust mechanism to improve language grounding in agent's action space**

LMC: Proposed Language Guided Meta-Controller



New Objective: Auxiliary Reasoning Loss

Predict a sequence of action types to improve ***"conceptual grounding"***

→ add a **linear layer** over the **multimodal transformer** to predict a sequence of action types

→ use a cross-entropy loss to train the auxiliary module

$$\hat{m}_t = f\left(x_{1:L}, v_{1:t}, \hat{a}_{1:t-1}\right) \quad \mathcal{L}_{aux} = \sum_{t=1}^{T} m_t \, log(\hat{m}_t)$$

Self Monitoring Agent: Progress Monitoring Loss [1]

→ Helps agent learn the utility of each state in the process of completing the overall task

$$p_t = f\left(x_{1:L}, v_{1:t}, \hat{a}_{1:t-1}\right) \quad \mathcal{L}_{pm} = \sum_{t=1}^{T} \left(y_t - p_t\right)^2$$

[1] Episodic Transformer for Vision-and-Language Navigation

Experimental Results

Model	Seen SR [TLW]	Seen GC [TLW]	Unseen SR [TLW]	Unseen GC [TLW]
Random	0.82 [0.62]	0.75 [0.43]	1.34 [0.43]	0.41 [0.07]
Lang	0.99 [0.28]	1.04 [0.29]	2.36 [0.23]	0.78 [0.29]
Vision	5.1 [1.15]	6.96 [1.76]	3.89 [0.61]	3.56 [0.73]
E.T.	9.5 [2.8]	10.0 [7.5]	7.6 [2.2]	9.1 [7.3]
+LMC (Ours)	18.55 [5.6]	19.0 [12.1]	12.5 [3.8]	12.0 [11.5]
+Aux	10.9 [2.6]	11.6 [8.0]	10.7 [2.8]	11.0 [10.4]
+PM	8.2 [3.6]	9.5 [8.1]	. 10.5 [3.6]	10.4 [11.1]

Evaluation Metrics:

- **Success Rate (SR):** 1 if all expected state changes are completed by the agent, else 0.
- **Goal-Condition Success (GC):** Fraction of all expected state changes completed by the agent. Averaged over all trajectories
- **Trajectory Weighted Metrics (TLW):** Evaluation metrics weighted by the trajectory length of the agent during the instance.

Summary

- We explore the problem of embodied instruction following while learning from human-human dialogues.
- We propose a language guided meta-controller that enables a more robust language grounding in the agents' action space.
- Our meta-controller leads to an **absolute improvement of 9% in success rate over seen environments, and 5% over unseen environments** in the validation set.