Review3: Habitat: A Platform for Embodied Research

Vivswan Shitole

CS539 Embodied AI

1 Summary

This paper presents a novel platform called "Habitat" for research in Embodied AI. The main contribution of this platform is its flexibility, modularity, reconfigurability and compatibility which can trivialize many experiments in the Embodied AI setting, which otherwise would require a lot of engineering effort to setup, run, validate and recreate. More concretely, the Habitat platform has a multi-layer stacked architecture with two major modules: (1) **Habitat-Sim**: its a flexible, high-performance 3D simulator responsible to simulate the environment by loading 3D scenes into a standardized scene graph representation. The hierarchical scene graph enables support for multiple 3D datasets (both synthetic and real-world based). It supports reconfiguring agents with multiple sensors, simulating agent motion and returning sensory data from an agent's sensor suite. It achieves a rendering rate of over 10,000 fps multi-process on a single GPU, which is orders of magnitude faster than the closest simulator. (2) **Habitat-API**: its a modular high-level library for end-to-end development of emnbodied-AI algorithms, defining embodied-AI tasks (navigation, instruction following, question-answering) and setting the training methodologies (imitation learning, reinforcement learning or classic SLAM). Such modularity trivializes conducting cross-dataset generalization experiments with different sensor configurations and different combinations of baseline learning algorithms.
To test the utility of the Habitat platform, experiments are carried out to test for generalization of goal-directed visual navigation agents between datasets of different environments. The goal is specified relative to the agent's initial position. The goal specification is static but the relative position can be continuously updatted via an ideal GPS (but no compass) that the agent is equipped with. The agent is embodied as a cylinder primitive shape with an action space of 4 actions (move forward, turn left, turn right, stop). The agent navigates in a continuous space supporting collisions and partial steps. The sensory input configurations yield 4 types of agents (blind, RGB, Depth and RGB + Depth), which are evaluated over 3 datasets (SUNCG, Matterport3D and Gibson). The agent's starting position and orientation are sampled uniformly at random from all navigable positions. One episode is constituted by 500 actions. The evaluation metric is "success weighted by path length" (SPL) = $l/max(p,l)$, where, l is geodesic distance of the shortest path and p is distance traversed by the agent. The baseline learning algorithms employed are random, forward-only, goal-follower, PPO and SLAM.
The key results obtained include: (1) learning outperforms SLAM when scaled over orders of magnitude more experience. (2) RGB agents don't significantly outperform Blind agents. (3) RL (PPO) + Depth agent performs the best across all datasets. (4) Average number of collisions incurred follow the ordering Blind ¿ RGB ¿ Depth. (5) All agents suffer a drop in performance when trained on one dataset and tested on other. Depth agent generalizes the best across datasets. (6) Most agents achieve higher performance over Gibson dataset than SUNCG or Matterport, implying Gibson dataset to be the simplest. Counter-intuitively, agents trained on Gibson consistently outperform their counterparts trained on Metterport3D and SUNCG, even when evaluated on Metterport3D and SUNCG.

2 Strengths

The platform introduced is not a minor improvement, its order of magnitudes better than other existing platforms, particularly in its rendering FPS, modularity, flexibility and compatibility with other datasets. It made possible the first cross-dataset generalization experiment in the embodied setting. The counterintuitive results obtained are a result of extensive experience scaling by virtue of the simulator's speed. The reconfigurability of the platform allowed for agressive rejection sampling of simpler environment configurations for the specified task. Generalization across different agent types over different environments lays the groundwork for simulation-to-reality transfer experiments.

3 Weaknesses

Too much experience scaling may be counter-productive since even a simple agent can perform orders of magnitude better when given something like a multi-month experience. The goal follower agent is specified to align itself when its not facing the goal (off the axis by more than 15 degrees). Its not clear how the misalignment is detected when the agent is not equipped with a compass. No ablation studies are conducted even though the platform is specified to support noisy sensors and ablations of agent's actions and actuators.

4 Reflections

The Habitat platform is a significant contribution to the arsenal of Embodied-AI Testbeds. As demonstrated in the paper, it can enable experiments that were too complex to conduct, leading to novel findings thus advancing research. I wonder if such a modular, flexible, compatible and reproducible testing framework can be created for the RL domain in general. Current testbeds (ALE, VizDoom, DeepMind Lab) are not easily reconfigurable and definitely don't help reproducibility of results.

5 Most Interesting Thought

The Habitat challenge is an interesting shift from the traditional "internet AI" challenges where participants train and predict on a static dataset. On the other hand, Habitat challenge for embodied-AI requires the participants to train on different environment settings which act as oracles generating dynamic dataset. Moreover, it requires the participants to upload their methodologies / solution code rather than just predictions.