22 December 2024 · CC-BY 4.0

Peer Review

Review of: "CityWalker: Learning Embodied Urban Navigation from Web-Scale Videos"

Kongwah Wan[1]

[1] Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

This is a review of the paper "CityWalker: Learning Embodied Urban Navigation from Web-Scale Videos"

This paper proposes learning human-like urban navigation by training agents on thousands of hours of web-sourced in-the-wild ego-motion videos of walking and driving. Pseudo action labels, represented by relative poses between frames, are extracted from these videos using Visual Odometry and fed into a data processing compute pipeline for large-scale imitation learning. During an online deployment, the task is cast as a point-goal navigation problem, wherein the agent receives as input camera images, a current GPS location, and a sub-goal waypoint coordinate. The agent augments the inputs with past observations and maps the combined input to an action like the typical cmd_vel. For fine-tuning and testing, 15 hours of tele-operation data across various unseen areas are allocated, with 6 hours for fine-tuning and 9 hours for testing.

The paper appears to continue the line of work of GNM [14] and ViNT [44] in using imitation learning to develop navigation policies in dynamic unseen environments by mapping images to actions. While GNM and ViNT use image-goal as target locations and GPS localization only during training, the current paper uses point-goal and GPS for localization. If GPS is used, the question is whether these GPS signals are the predominant inputs to enable the agent to successfully reach the target location.

The authors mention that another key goal of the work is to "focus on learning the rules and norms essential for urban environments, which is crucial for enabling applications such as delivery robots and robo-taxis." However, there appears to be little evaluation of how such rules and norms are adequately learned.

doi.org/10.32388/35WSAI

In the experimental section “Real-World Deployment,” the authors describe categorizing the test trajectories into FORWARD, LEFT, and RIGHT cases. The authors then report "robust performance" in both forward and turns, which “highlights its ability to manage dynamic and varied urban scenarios effectively, allowing it to be more reliably used in real-world environments where frequent and precise directional changes are essential” (page 6). Question: Are the test trajectories categorized based on whether the robot makes any left or right turn? For example, even if a test trajectory is a straight forward path, while navigating this trajectory, if the robot performs right/left turns to avoid obstacles, is this trajectory categorized as LEFT/RIGHT?

Overall, this is a good effort that continues the push to develop robust navigation policies in unseen and dynamic urban environments.

Declarations

Potential competing interests: No potential competing interests to declare.