Goal-Oriented Robot Navigation to Arbitrary Objects

A person slips and calls for a first-aid kit. A truly useful household robot must fetch it, even with no prior experience doing so. In these time-sensitive, high-stakes applications, it is desirable to convey only the goal instead of low-level actions as in my prior work [1]. The robot should understand the new task and act autonomously. These capabilities are beyond current techniques, which often train robots for specific problems. This motivates developing robots that complete new tasks without additional training to address the unstructured and unpredictable nature of everyday life.

I will **move towards general purpose robots** fetching objects by **developing robots that navigate to arbitrary objects**, which is a critical step. Here, an active agent must explore based on what it sees to find target objects with no prior knowledge of the test object categories, like first-aid kits. This departs from prior work [2], which assumes fixed common categories. The ability to understand arbitrary images is helpful to address this gap. Recent work points to the success of training large neural networks on internet data. For example, a recently published algorithm called CLIP [3], is trained to map internet images and their corresponding captions to the same numerical feature representation. After training, it is possible to specify arbitrary prompts like “a red inhaler” that will have similar representations to images of red inhalers. Hence, it becomes possible to check for arbitrary concepts in images using the appropriate prompt. While extremely powerful, my initial experiments suggest CLIP cares more about *what* is in an image than *where* it is. To determine this, I designed prompts with actionable words (e.g., “inhaler to the right” v.s. “inhaler to the left”) and ran experiments to see if CLIP could correctly match against validation images. I found the accuracy was no better than random chance at ~50%. These limitations imply that improvements are needed for such algorithms to be useful in robot navigation settings where right and left are important.

Research Proposal

My goal is to enable a robot to find a desired object specified by a human in text, like a first-aid kit. The robot should explore rooms autonomously until it finds the desired object. More formally, given a text caption of a target object and visual observations, the task is to find a sequence of actions from the current position to the target. I will consider a robot with a wide angle camera sensor and three navigation actions: <*turn left*>, <*move forward*>, and <*turn right*>.

Aim 1: Navigating to observed objects

Motivated by peoples’ inclination to **look around with their eyes before moving**, my key insight is to link actions to regions of visual observations. To connect vision and actions, I will take left, center, and right image crops of an agent’s view. These crops are akin to a human looking from left to right to survey a room. If a human notices the target object to the left, then they know to move in that direction. Similarly, for each crop, I will leverage CLIP to check for the target object specified by user-generated text input. If the object is found in the crop with high confidence, the agent will take the action to move in this direction. The simplicity of this novel method is also one of its greatest strengths. It will make it easy for other roboticists to incorporate my method into their pipelines.

I will first set up experiments in a simplified setting, where an agent must reach a fully observable target with no obstacles. To do so, I will use the THOR household robot simulation environment [4]. To measure effectiveness of the cropping strategy, I will calculate success by **checking if the agent gets within 0.5m of target objects**. As this is a simpler version of the ultimate task, I will consider it complete when reaching the high benchmark of **90% success**. When this is the case, I will add distractor objects to verify the agent can move to the correct object among many choices. I will scale the complexity of my approach to address failure cases if they arise. For far away objects, I may need to zoom into regions that would otherwise get lost. I will adjust the crops of the input images using a region proposal network, which is a type of neural network that extracts patches of images corresponding to objects. While slightly more computationally expensive, the modification will give areas to look at more closely.

Aim 2: Exploring rooms to find objects

Motivated by peoples' **moving around when searching for things**, I will train an exploration algorithm to address cases when the target object is not in view or is obstructed. Once there is a clear path to the object, the agent will use the crop module from Aim 1 to reach the goal. For exploration, I will use deep reinforcement learning techniques, where learning happens by trial and error. I will train the neural network to choose the best of the three actions in THOR. I will provide the network with rewards only when it observes new things and does not bump into obstacles. For metrics, I will **compute percentages** of the room observed and of actions that do not result in collisions. Similar investigations by colleague Dr. Kiana Ehsani suggest the method will transfer to the real world. I will train until I achieve **75% for both percentages**, which will indicate adequate exploration.

A tuned weighted-average over the action predictions from the crop and exploration modules will give a baseline that looks and moves around. The agent should follow predictions from the module that is most confident to get the best of both worlds. To tune this approach, I will generate validation data in THOR and measure object navigation metrics, namely **success rate** and **success weighted by path length (SPL)**, which penalizes inefficient trajectories. To push the research further, I will iterate on strategies for combining the predictions from both modules to improve the performance metrics. For example, I can train another neural network that optimizes for downstream performance directly to replace the weighted-average operation. If this technique does not show significant improvement, it would still be a notable result, suggesting the simple baseline is hard to beat.

Aim 3: Real world experiments

To ensure my methodology is **applicable in real scenarios**, I will test my algorithms using a Fetch robot in my lab. I will run the same pipelines in both simulation and the real world. Because CLIP has been demonstrated to work on synthetic and real images, I suspect success rate should be similar. To verify, I will create validation datasets, plot simulated success on the x-axis, and real success on the y-axis. I expect to see a **linear trend** with a line of best fit close to $y = x$. If this is not the case, I will tune simulation parameters following best-practices reported in the literature [5]. I will conduct a **user study** to expose if the algorithm can discover objects specified by people from all walks of life.

Intellectual Merit

The proposed research provides a path towards grounding internet-trained models in robot embodiment, drawing on how humans behave for inspiration. Such innovation will allow roboticists to easily benefit from advances in such models. The project addresses scenarios where humans tell robots to find arbitrary things, which itself could inspire roboticists to think of creative alternatives to conventional task-specific training. I am uniquely positioned to conduct this research considering my background with simulators, neural networks, robot platforms, and conducting user studies.

Broader Impact

I will use this project as a platform to create teaching resources that are accessible to the community as addressed in the Personal Statement. This project also has long-term implications on changing the way that people live their everyday lives. Objects people care about in the real world are often very specific, like "a red inhaler" or "vial of insulin." This proposal makes significant progress in robots finding these objects, thereby combating bias around what objects are deemed important for algorithms to consider. This will make it possible to tailor robot capabilities to the individual, which make them even more useful in homes.

References

[1] S.Y. Gadre, et al. *End-User Robot Programming Using Mixed Reality*. ICRA 2019.; **[2]** M. Wortsman, et al. *Learning to Learn How to Learn: Self-Adaptive Visual Navigation Using Meta-Learning*. CVPR 2019.; **[3]** A. Radford, et al. *Learning Transferable Visual Models From Natural Language Supervision.* ICML 2021.; **[4]** E. Klove, et al. *AI2-THOR: An Interactive 3D Environment for Visual AI*. Pre-print 2017.; **[5]** A. Kadian, et al. *Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance?* RA-L 2020.