Teaching Artificial Intelligence Systems to effectively and adaptively understand the real world.

Anirudha Ramesh

Over the past decade, artificial intelligence has seen an explosive growth on numerous fronts, particularly in computer vision and NLP. Focusing on vision, most of this progress has however been in low level perception tasks, and our key weapon to aid generalization has just been to try to bring more data in-domain. The next steps in our journey relate to making our vision systems capable of higher level reasoning, and more adaptable. Higher level reasoning would allow our systems to naturally navigate the real world, and adaptability is the key pillar on which the house of artificial general intelligence will be built.

These two threads are in fact not disjoint, rather complement each other. A system capable of higher level reasoning should be able to reason and adapt to new tasks and environments more easily and effectively by relating its current capabilities to whatever new is put in front of it. Humans can adapt to a wide range of environments and tasks with very minimal supervision / guidance once they have developed a good level of reasoning, provided they are given a resource to extract target-domain information/experience. For example, a person trained to drive a specific car in Pittsburgh, can under most circumstances, drive most cars, in most places around the world. In fact, having a driver's license is usually the only requirement to rent out a jet-ski in most places around the US.

Bringing these together, the longer term goals of my research include creating adaptable vision systems that can leverage their higher level reasoning to perform varied tasks in varied environments. This, in turn, would lead to robots capable of working in a magnitude of domains, performing a wide range of tasks without large scale human supervision, or human aided system changes. To solve this, we need to :-

1) Create **visual representations that inherently capture a greater amount of semantic information** about objects and scenes, as well as information about how objects in a scene relate to each other, and the world.
2) Create foundational **models which can adapt to similar tasks, and extend their capacity in the source task**, with minimal guidance.
3) Create foundational **models which can work in new environments**, i.e. domains, with minimal guidance, by leveraging understanding of whatever domains it has already seen, and in extension, understanding where this new domain is different.
4) Connect our representations, and models, to real robots. Doing this would require creating and working in **spatially meaningful representations, and navigating the real world**.

Over the last few years, I’ve had the opportunity to work on all these different problems, in specific flavors. I summarize some of my contributions in the following paragraphs, in chronological order.

Navigating the real world

Safely and correctly moving around in the real world has always been one of the key problems in robotics. Autonomous driving, particularly in on road environments, has been one of the domains that has generated the most interest amongst researchers. My contributions to this problem are two fold.

First and foremost, I developed a novel method to bring monocular multibody SLAM frameworks into a uniform metric scale [1], thus solving the fundamentally ill-posed scale problem in a monocular setting. This is particularly useful in the miniaturization of autonomous agents, where we might not have the capacity to equip and process information from multiple cameras. Solving this scale problem helps us better plan our paths, and control inputs, thus making it possible to navigate safely and effectively. We extend our work by introducing a state-of-the-art method to perform SLAM and provide a metric representation in a Bird's Eye View suitable for downstream tasks like planning and prediction in [2].

We also tackled the highly challenging problem of navigating in a world with other navigating agents whose behavior/intents are unclear to us in [3]. We propose a novel method for reactive multiagent collision avoidance by characterizing the longitudinal and lateral intent uncertainty along a trajectory as a closed-form probability density function.

Bringing both these threads together, our work has helped push the boundaries of the abilities of autonomous agents to move around the real world.

Few Shot Learning

In our path to making robots and ML systems adaptable over various tasks, our first order of business is making our system adaptable within the same task. To this end, we tackle the problem of Few Shot Segmentation [4], wherein we identify key biases (Saliency Bias and Class Negative Bias) that were a part of most existing methods, and all training and evaluation paradigms on standardly used datasets. We proposed a general purpose algorithm to remove some of these biases, and proposed a new tiered dataset for a more meaningful evaluation of FSS algorithms. Our work has been key in improving segmentation of new objects in Adobe's automatic (particularly large-scale) photo-editing capabilities.

Semantically Enhanced Visual Representations

Human's have a great understanding of object inter-relations with both the environment, and other objects in the scene/world. This comes from a deeper understanding of the world which most vision systems still do not possess. With this understanding, most human decisions are predictive in nature, as can be confirmed by years of research in predictive coding theory. This relates to both estimations involving physical laws such as gravity, as well as, decision making in the world (eg :- where would you look for your keys first in case you lose it?). In most object localization tasks we rely on our predictive abilities first, followed by detection capabilities. With this intuition, we develop a pre-training task to teach our detection models to learn to predict where objects are likely to go into a scene, and in doing so we improve its performance in detection, particularly recall [5]. Our work raises many questions, still unanswered, on the role and necessity of building vision systems with predictive capabilities encoded in them.

Adapting and Generalizing to arbitrary domains

Generalization and competitive performance in out-of-domain distributions is a key challenge on the path to seeing robots operate in less-constrained real world environments. There are countless ways to perceive the world around us, and our robots need not be restricted to just doing so on standard RGB images, which are usually from datasets that don't always capture all the diversity in the world. Introduction of new-modalities can extend the operating capabilities of our robots, (eg :- how would you see in a cave? underwater? In the dark?), but training ML systems on arbitrary modalities is very challenging owing to the lack of data in these domains. At this end, we propose a new view on this problem by bridging the gap between domain generalization and adaptation, and conduct a comprehensive study to understand the different aspects of our models that are necessary for adapting across different kinds of domains. Our work enabled us to extend the operating capacity of fleets of robots to work off-road, at-night, with only passive sensors.

Future Research Plans

My mid-term research goal is to bring together higher-level reasoning capabilities in vision / ML systems with adaptivity, in both the task space, and operating domains.

- A key direction for this is creating semantically enhanced visual representations that are specifically engineered so as to be easily adaptable. These are two directions that are seldom viewed in conjunction, but are actually key to enhanced visual intelligence, particularly in robotics.
- Given the plethora of headway we've seemingly made in NLP, and seem on the brink of achieving in vision, in semantic understanding of the visual world, for robotics it is imperative that we connect this to the physical world. Given the seemingly sophisticated nature of these reasoning systems, it's crucial that we explore and exploit their capabilities to guide generalizability in robotics.
- I would also like to explore the potential and try to instantiate neural communication across humans and robotic systems first, and other humans next. Humans after all, are the most important parts of the world most robots will inhabit.

References

1) Multi-object Monocular SLAM for Dynamic Environments : https://arxiv.org/abs/2002.03528
2) BirdSLAM: Monocular Multibody SLAM in Bird's-Eye View : https://arxiv.org/abs/2011.07613
3) Probabilistic Collision Avoidance for Multiple Robots : A closed form PDF approach : https://ieeexplore.ieee.org/document/9575767
4) What Ails One-Shot Segmentation : A data perspective : https://openreview.net/forum?id=BlcUQYxknbX
5) Learning to Detect by Learning to Predict : https://anirudharamesh.github.io/posts/2012/08/Learning-to-Detect-by-Learning-to-Predict/