THE INGREDIENTS OF REAL-WORLD ROBOTIC REINFORCEMENT LEARNING

Training robots by having them operate in the 'real' world.

Based on the paper: https://openreview.net/pdf?id=rJe2syrtvS

Gene Olafsen

ACKNOWLEDGEMENT

- See paper: https://openreview.net/pdf?id=rJe2syrtvS

- Henry Zhu∗1, Justin Yu∗1, Abhishek Gupta∗1, Dhruv Shah1,Kristian Hartikainen2, Avi Singh1, Vikash Kumar3, Sergey Levine1
- 1 University of California, Berkeley 2 University of Oxford 3 University of Washington

THE PROMISE

- Reinforcement learning, in principle can enable autonomous systems, such as robots, to acquire a large repertoire of skills automatically.
- Reinforcement learning can enable such systems to continuously improve the proficiency of their skills from experience.

THE REALITY

- Realizing this in reality has proven challenging: even with reinforcement learning methods that can acquire complex behaviors from high-dimensional low-level observations, such as images, the assumptions of the reinforcement learning problem setting do not fit cleanly into the constraints of the real world.

RL AND REAL WORLD TRAINING

- Most successful robotic learning experiments have employed various kinds of environmental instrumentation in order to:
  - Define reward functions
  - Reset between trials
  - Obtain ground truth state

CHANGE IS NEEDED

- In order to practically and scalably deploy autonomous learning systems that improve continuously through real-world operation, we must lift these limitations and design algorithms that can learn under the constraints of real-world environments.

PROPOSED CHANGES

- The authors of the paper propose that overcoming these challenges in a scalable way requires designing robotic systems that possess three capabilities:
  - Able to learn from their own raw sensory inputs
  - Assign rewards to their own trials without hand-designed perception systems or instrumentation
  - Learn continuously in non-episodic settings without requiring human intervention to manually reset the environment

ILLUSTRATION

Figure 1: Illustration of our proposed instrumentation-free system requiring minimal human engineering. Human intervention is only required in the goal collection phase (1). The robot is left to train unattended (2) during the learning phase and can be evaluated from arbitrary initial states at the end of training (3). We show sample goal and intermediate images from the training process of a real hardware system

MARKOV

- The standard reinforcement learning paradigm assumes that the controlled system is represented as a Markov decision process with a state space S, action space A, unknown transition dynamics T , unknown reward function R, and a (typically) episodic initial state distribution ρ. The goal is to learn a policy that maximizes the expected sum of rewards via interactions with the environment.

MEET THE REAL WORLD

- Although this formalism is simple and concise, it does not capture all of the complexities of real world robotic learning problems.
- For a robotic system is to learn continuously and autonomously in the real world, we must ensure that it can learn under the actual conditions that are imposed by the real world.

FROM THE 'MDP' TO THE REAL WORLD

- 1) All of the information necessary for learning must be obtained from the robot's own sensors. This includes information about the state and necessitates that the policy must be learned from high-dimensional and low-level sensory observations, such as camera images.
- 2) The robot must also obtain the reward signal itself from its own sensor readings. This is exceptionally difficult for all but the simplest tasks (e.g., reward functions that depend on interactions with specific objects require perceiving those objects explicitly).
- 3) The system must be able to learn without access to episodic resets. A setup with explicit resets quickly becomes impractical in open-world settings, due to the requirement for significant human engineering of the environment, or direct human intervention during learning.

PRIOR APPROACHES

Figure 2: We draw a comparison between current real world learning systems which rely on instrumentation versus a system that learns in an environment more representative of the real world, free of instrumentation. While all three prior works utilize instrumentation for resets, state estimation, and reward, the motion capture system of Gupta et al. (2016), sensor attached to the door in Zhu et al. (2019), and auxiliary robot which picks up fallen balls in Nagabandi et al. (2019) are good examples of engineered state estimation, reward estimation, and reset mechanisms respectively.

LEARNING FROM RAW SENSORY INPUT

- Require the robotic systems to be able to learn from their own raw sensory observations. Typically, these sensory observations are raw camera images from a camera mounted on the robot, as well as proprioceptive sensory inputs such as the joint angles.

REWARD FUNCTIONS WITHOUT REWARD ENGINEERING

- In the real world, the robot must obtain the reward signal itself from its own sensor readings.
- One candidate is for a user to specify intended behavior beforehand through examples of desired outcomes (i.e., images).
- The algorithm can then assign itself rewards based on a measure of how well it is accomplishing the specified goals, with no additional human supervision.
- This approach can scale well in principle, since it requires minimal human engineering, and goal images are easy to provide.

LEARNING WITHOUT RESETS

- Natural open-world settings do not provide any such reset mechanism, and in order to enable scalable and autonomous real-world learning we need systems that do not require an episodic formulation of the learning problem.
- To devise a system that requires minimal human engineering for providing rewards, we must use algorithms that are able to assign themselves rewards, using learned models that operate on the same raw sensory inputs as the policy

THE CHALLENGES OF REAL-WORLD RL

- The researchers built a system largely defined by the principles identified in this paper.
- However, when utilized for robotic learning problems, they found this basic design to be largely ineffective.
- They presented results for a simulated robotic manipulation task that requires repositioning a free-floating objects with a threefingered robotic hand.
- The goal in this task is to reposition the object to a target pose from any initial pose in the arena.

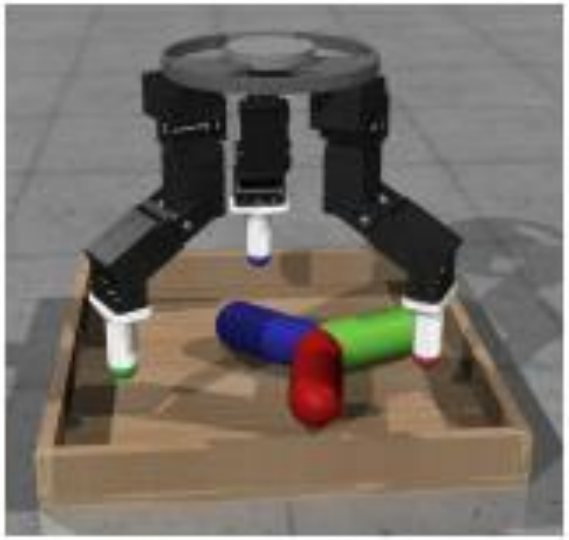

Figure 3: Our object repositioning task. The goal is to move the object from any starting configuration to a particular goal position and orientation.

FAILURE TO MAKE PROGRESS

- When the system is instantiated with vision-based soft actor-critic, rewards from goal images using VICE, and run without episodic resets, we see that the algorithm fails to make progress.
- Although it might appear that this setup fits within the assumptions of all of the components that are used, the complete system is ineffective.

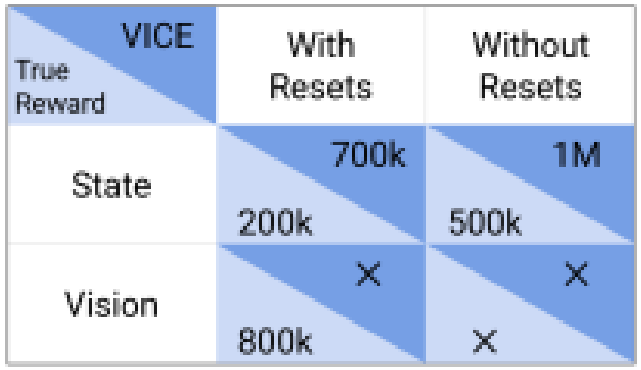

Figure 4: We report the approximate number of samples needed for a policy learned with a prior *off-policy RL algorithm (SAC)* to achieve average training performance of less than 0.15 in pose distance (defined in Appendix C.1.3) across 3 seeds on the repositioning task. We compare training performance after varying three axes: ground truth rewards vs. learned rewards, with vs. without episodic resets, low-level state vs. images as inputs. We observe learning without resets is harder than with resets and is much harder when combined with visual inputs.

EXPERIMENT

- To investigate the issue concerning poor training results, the team performed experiments investigating the combination of the three main ingredients:
  - Varying observation type (visual vs. low-dimensional state),
  - Reward structure (VICE vs. hand-defined rewards that utilize ground-truth object state),
  - The ability to reset (episodic resets vs. reset-free, non-episodic learning).

RESET FREE CHALLENGE

- The results show that learning with resets achieves high training time reward from both vision and state, while reset-free only achieves high training time reward with low-dimensional state.
- Second, the policy is able to pass the threshold for training time reward in a non-episodic setting when learning from low-dimensional state, but it is not able to do the same using image observations.
- This suggests that combining the reset-free learning problem with visual observations makes it significantly more challenging.