GPU-Accelerated Deep Reinforcement Learning Simulation For Vision-Based Robotic Manipulation

Kihiro Bando, Kingway Liang, Victor Zhang {bandok, kingway, zhangvwk}@stanford.edu

Summary

Context

- Deep reinforcement learning (DRL) is a promising approach for complex robotic tasks but requires large amount of data [2].
- Data collection via simulations has become increasingly popular thanks to research in sim-to-real transfer.

Approach

- FleX GPU-based simulator has been shown to provide substantial speed-up while requiring less hardware resources [1].
- Goal: demonstrate the feasibility of using FleX as an environment for robotic DRL.

RL formulation

- **State space**: absolute and relative positions of the end–effector ($\mathcal{S}_{low}$); 84x84x4 RGBD image and absolute position of the end-effector ($\mathcal{S}_{high}$).
- **Action space $\mathcal{A}$**: translational velocity.
- **Reward**: L1 distance to the target.
- **Policy π_θ**: normal distribution with learned mean and log standard deviation.

Algorithm

Proximal Policy Optimization (PPO) [3]

- Variant of on-policy policy gradient method.
- Loss:

$$L(\theta) = \mathbb{E}\left[\min\left(r_t(\theta),\ \mathrm{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\right)\hat{A}_t - \left(V_\theta(s_t) - V_{\mathrm{target}}(s_t)\right)^2\right]$$

- Alternates between rollouts of data collection and optimization steps.

References

[1] Liang *et al* (2018)

[2] Kalashnikov *et al*, (2018)

[3] Schulman *et al* (2017)

Architecture

With low-dimensional observation space

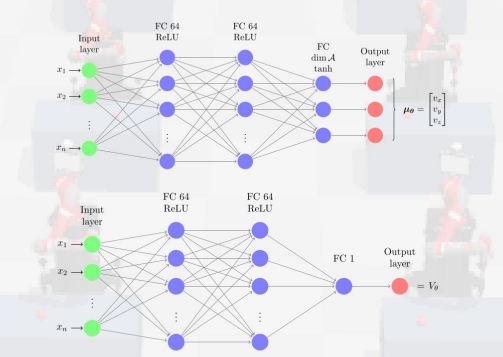

Figure 1: Actor-critic network architecture

With high-dimensional observation space

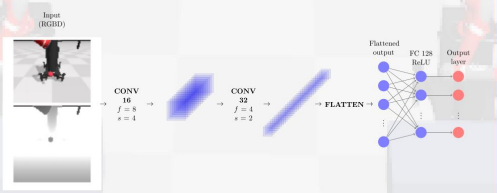

Figure 2: CNN image encoder

- The output is then concatenated with the position of the end-effector and fed into the actor-critic networks (top).

Results

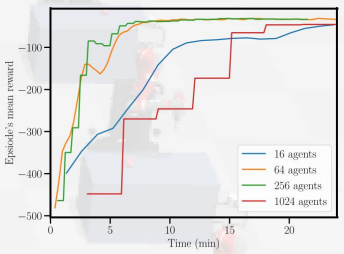

Figure 3: Average return over all episodes so far vs wall-clock time

Qualitative analysis: the scaling shows an optimal balance in order to collect full episodes while minimizing computational time. The results are sensitive to reward and state formulations. The agents do not learn correctly with the L2 distance or without the information about the relative position. Coordination with the environment is important. The learning is improved when not resetting the agent upon reaching the target.

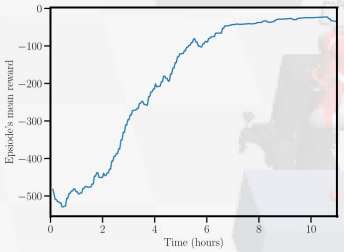

Figure 4: Average return with vision input

Vision-based learning is a much harder task with many more weights. The algorithm is also more prone to instabilities. Batch normalization for instance is necessary. The optimization of the data collection across many agents is a challenge.