IBN-ML-STUDY WEEK14 PAPER REVIEW

Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates

by Shixiang Gu, Ethan Holly, Timothy Lillicrap and Sergey Levine

CONCEPTS

- Introduction
- Related work
- Background
- Asynchronous training of normalized advantage functions
- Simulated experiments
- Real-world experiments
- Discussion and future work
- Personal Opinion

INTRODUCTION

- "Reinforcement learning methods have been applied to range of robotic control tasks from locomotion, to manipulation and autonomous vehicle control"
- Limitation of reinforcement learning
  - policy or value function should be defined -> related to hardware(limit) and training time
  - needed initialize policy and safety concerns -> human demonstrations

- In this work, using off-policy training of deep Q-functions
  - not necessarily user-provided description or policies

INTRODUCTION

- Off-policy deep Q-function
  - Deep Deterministic Policy Gradient algorithm (DDPG)
  - Normalized Advatage Function algorithm (NAF)

- Reduce training times by parallelizing across multiple robotic platforms
  - random-target reaching
  - door pushing
  - door pulling
  - pick-and-place
- Without any human-provided examples for initialization (from scratch)

RELATED WORK

- Other application of reinforcement learning
  - low-dimensional policy representations (mentioned previous slide) -> under a hundred parameters
  - recent research deals with 7 DoF arms, continuos control of high-dimensional system

- Model-based vs. Model-free
  - model-based
    - successful on a range of real-world tasks
    - difficult in domains with severe discontinuities in the dynamics and reward function
  - model-free
    - DDPG, NAF, NFQCA
- Difference between similar research
  - minimize the training time when training on real physical robots
  - from scratch, without initializing the user demonstrations

BACKGROUND

- Explanation of fundamental reinforcement learning
- No information of NAF, DDPG but references

ASYNCHRONOUS TRAINING OF NORMALIZED ADVANTAGE FUNCTIONS

- Asynchronous Learning
  - Learner thread = trainer thread : 1
  - Experience collecting worker thread : n
  - worker thread send observation, action and reward for each time step
  - Each robots could run in real time, without experiencing delays due to the time complex

- Safety Constraints
  - joint's maximum velocity, position limits
  - bounding sphere, using Forward Kinematics
  - more required for safety

- Network Architectures
  - state : target postion
  - reaching task : end-effector

SIMULATED EXPERIMENTS

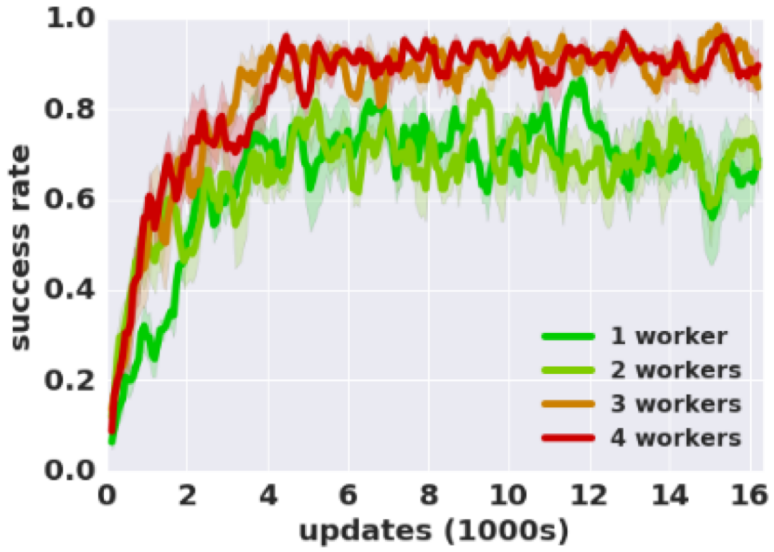
- Using 7-DoF arm and JACO arm
  - 7-DoF : random-target reaching, door pushing, door pulling
  - JACO : pick & place (JACO has three fingers, 9 DoF)
- Simulation Tasks & Neural Network Policy Representations

	Max. success rate (%)			Episodes to 100% success (1000s)		
	DDPG	Lin-NAF	NAF	DDPG	Lin-NAF	NAF
Reach	100±0	100±0	100±0	3.2±0.7	8±3	3.6±1.0
Door Pull	100± 0	5 ± 6	100± 0	10±8	N/A	6±3
Door Push	100±0	40± 10	100± 0	3.1± 1.0	N/A	4.2± 1.0
Pick & Place	100±0	100±0	100±0	4.4± 0.6	12± 3	2.9±0.9

Fig. 4: The table summarizes the performances of DDPG, Linear-NAF, and NAF across four tasks. Note that the linear model learns the perfect reaching and pick & place policies given enough time, but fails to learn either of the door tasks.

SIMULATED EXPERIMENTS

- Asynchronous Training – communication of devices



(a) Reaching

(b) Door Pushing

https://sites.google.com/site/deeproboticmanipulation/

REAL-WORLD EXPERIMENTS

- big difference with simulated work : friction, tight joint limits etc.
- reduced the performance
  - relax the definition of a positive episode count
  - bigger target region
- Random Target Reaching, Door Opening
  - because of geometry of robot, could not test door pushing task
  - no pick-and-place task in real-world
  - operating frequency of learning and training robots is affected to success rate

DISCUSSION AND FUTURE WORK

- Some limits about human demonstration or initial constraints

- Each robots collected data from different door or conditions
  - in this paper, robots work at same environment i.e. same size of doors
  - what if different shape or environment doors?

PERSONAL OPINION

- 방법과 결과를 같이 적어서 읽기 힘들었다.
- 구체적인 방법을 써 놓지 않고 **reference**로 적어 놨기에 추가적인 논문 탐색이 필요하다. (**NAF, DDSP, locomotion** 적용)
- 오탈자가 존재
- 너무 광범위한 지식이 필요

  ex) device간의 통신 모듈 이해, 딥러닝, 로봇역학 등
- 너무 어려운 논문으로 시작한 것 같다.