Hypernetwork-PPO for Continual Reinforcement Learning

Final Presentation

Philemon Schöpf

Supervisors: Sayantan Auddy, Jakob Hollenstein, Antonio Rodriguez-Sanchez

2022-09-29

Continual Reinforcement Learning

- Reinforcement Learning
  - Learn by interacting with an environment + getting rewards
  - Unsupervised - no training data, just an environment

Continual Reinforcement Learning

- Reinforcement Learning
  - Learn by interacting with an environment + getting rewards
  - Unsupervised - no training data, just an environment
- Continual
  - Learn multiple tasks sequentially
  - Cannot revisit old environment when learning new tasks
  - Do not forget old skills
  - Still a major issue in machine learning[2]

Proximal Policy Optimization

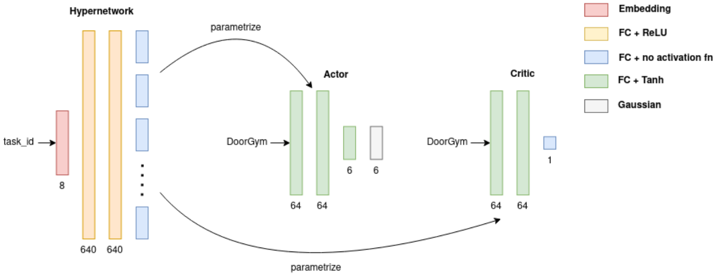
- On-line RL algorithm
- Objective is a "clipped" loss - discourages large, detrimental changes[3]

$$L_t^{clip}(\theta) = \mathbb{E}_t \left[min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t, clip \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1+\epsilon, 1-\epsilon \right) \hat{A}_t \right) \right]$$

- Additional loss components
  - state value
  - entropy bonus

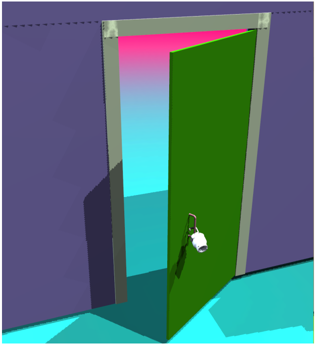
Hypernetworks

- Network that outputs a network[5]
- Task ID as input
- Target networks determine policy/dynamics
- Regularization on changes of outputs for old tasks



DoorGym

- Based on OpenAI Gym[1]
- Robot arms try to open doors
- Multiple handles, opening directions
- Our experiments
  - "Floating hook" robot
  - 6 different kinds of doors

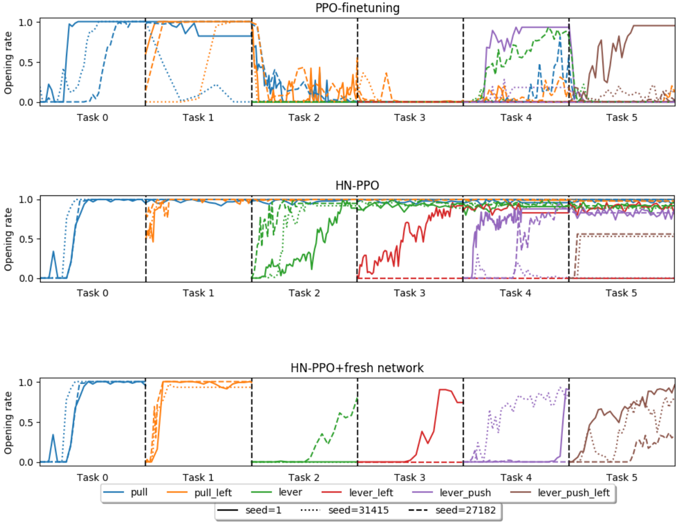


DoorGym world: `pull` handle, right hinge

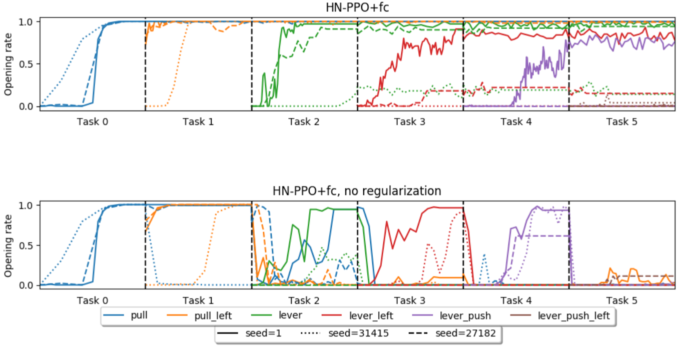
Experiments

- Baselines
  - PPO (pre-implemented in DoorGym)[4]
  - PPO-finetuning
  - HN-PPO with fresh networks for each task
- 2 hypernetwork architectures
  - HN-PPO
  - HN-PPO with fresh critic
- Ablation Study: HN-PPO without regularization

HN-PPO protects against catastrophic forgetting



HN regularization is required for CL performance



DoorGym demo

Conclusion

- HN-PPO is very effective against catastrophic forgetting
- Single-task success rate comparable to PPO
- Regularization crucial for HN-PPO's CL capability
- Limitations
  - Seed dependence
  - Checkpoint dependence

References

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba.
OpenAI Gym, 2016.

M. McCloskey and N. J. Cohen.
Catastrophic interference in connectionist networks: The sequential learning problem.
volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov.
Proximal policy optimization algorithms, 2017.

Y. Urakami, A. Hodgkinson, C. Carlin, R. Leu, L. Rigazio, and P. Abbeel.
DoorGym: A scalable door opening environment and baseline agent, 2019.

J. von Oswald, C. Henning, J. Sacramento, and B. F. Grewe.
Continual learning with hypernetworks.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.