Hypernetwork-PPO for Continual Reinforcement Learning

Final Presentation

Philemon Schöpf

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

2022-10-11

Continual Reinforcement Learning

- Reinforcement Learning
  - Learn by interacting with an environment + getting rewards
  - Training data collected from environment

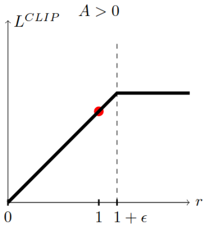
Continual Reinforcement Learning

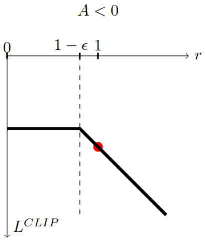
- Reinforcement Learning
  - Learn by interacting with an environment + getting rewards
  - Training data collected from environment
- Continual
  - Learn multiple tasks sequentially
  - Cannot revisit old environment when learning new tasks
  - Do not forget old skills
  - Still a major issue in machine learning

Proximal Policy Optimization

- On-policy, model free RL algorithm
- Objective is a "clipped" loss - discourages large, detrimental changes

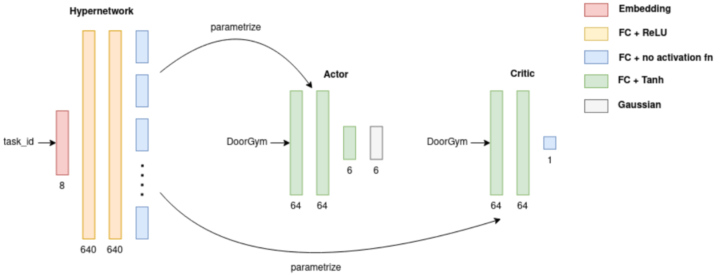
$$L_t^{clip}(\theta) = \mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t, clip\left(r_t(\theta), 1+\epsilon, 1-\epsilon\right)\hat{A}_t\right)\right]$$





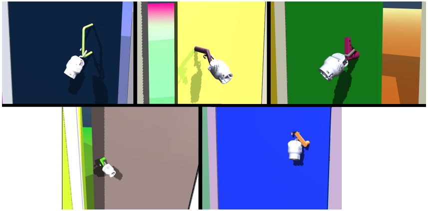
Hypernetworks

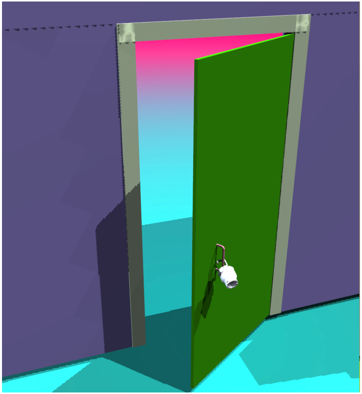
- Network that outputs a network
- Task ID as input
- Target networks determine policy/dynamics
- Regularization on changes of outputs for old tasks



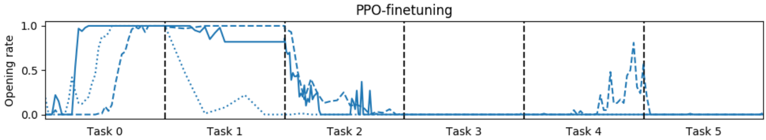
DoorGym

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

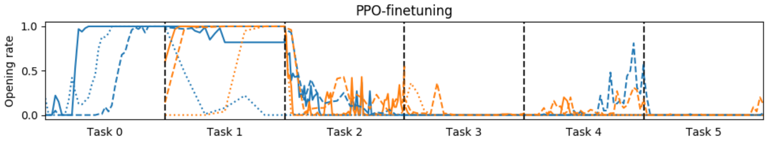




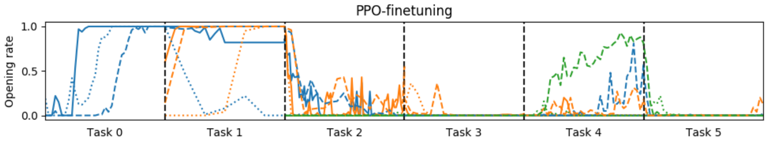
HN-PPO protects against catastrophic forgetting



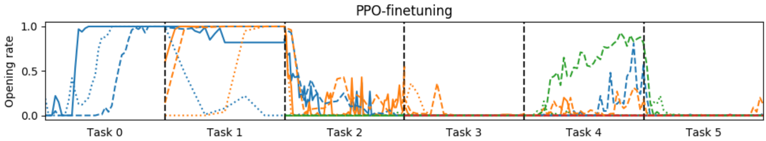
HN-PPO protects against catastrophic forgetting



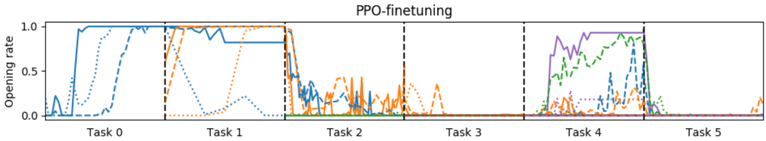
HN-PPO protects against catastrophic forgetting



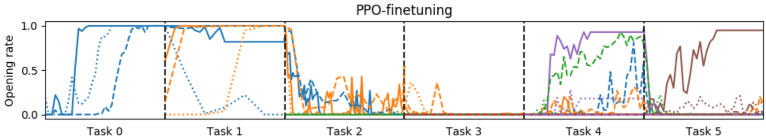
HN-PPO protects against catastrophic forgetting



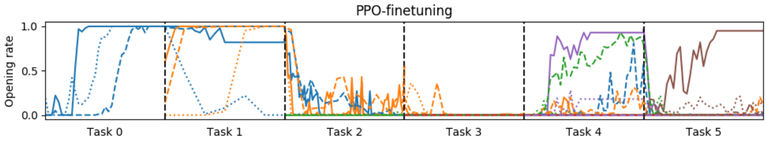
HN-PPO protects against catastrophic forgetting

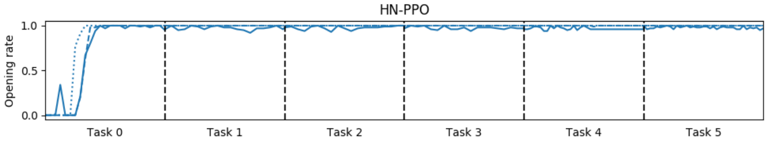


HN-PPO protects against catastrophic forgetting

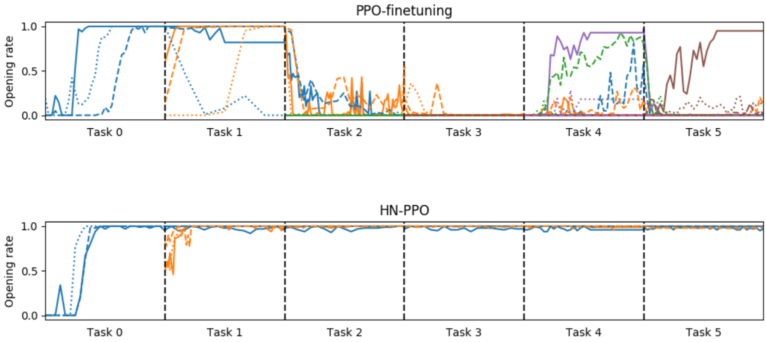


HN-PPO protects against catastrophic forgetting

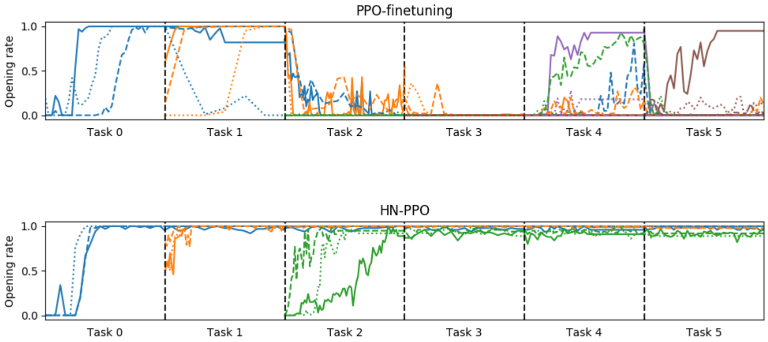




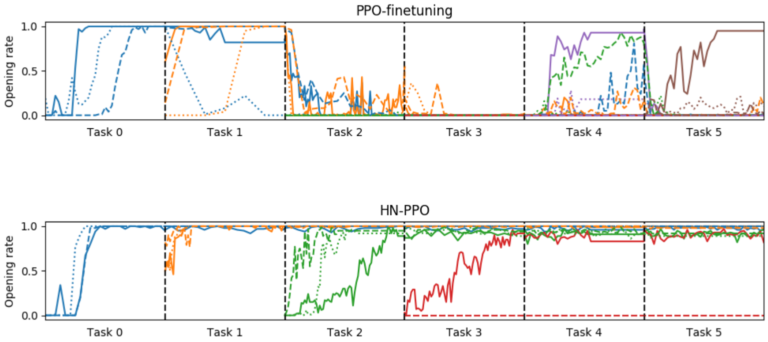
HN-PPO protects against catastrophic forgetting



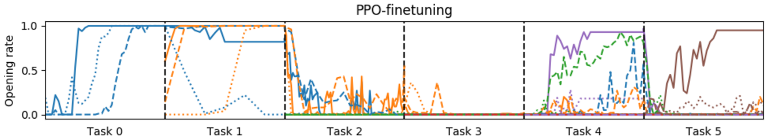
HN-PPO protects against catastrophic forgetting

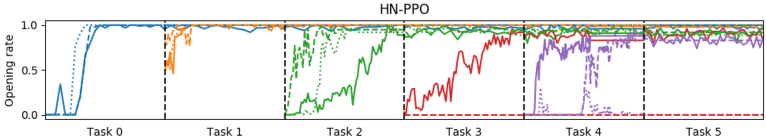


HN-PPO protects against catastrophic forgetting

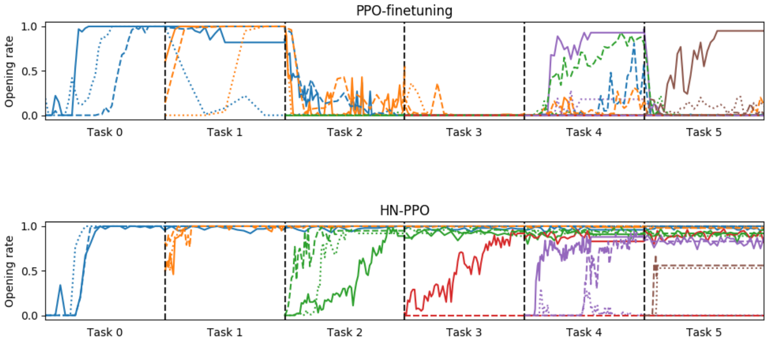


HN-PPO protects against catastrophic forgetting





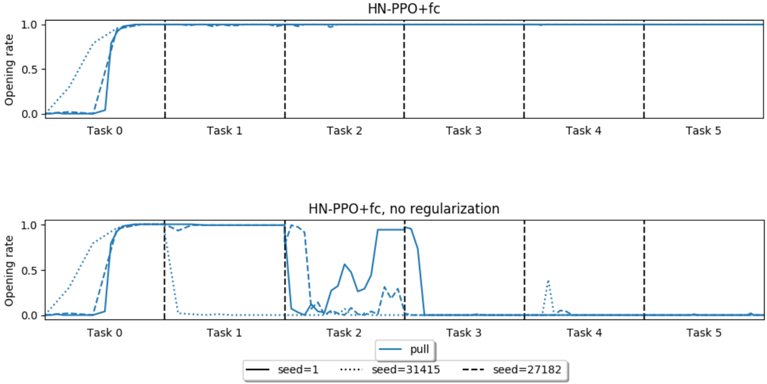
HN-PPO protects against catastrophic forgetting



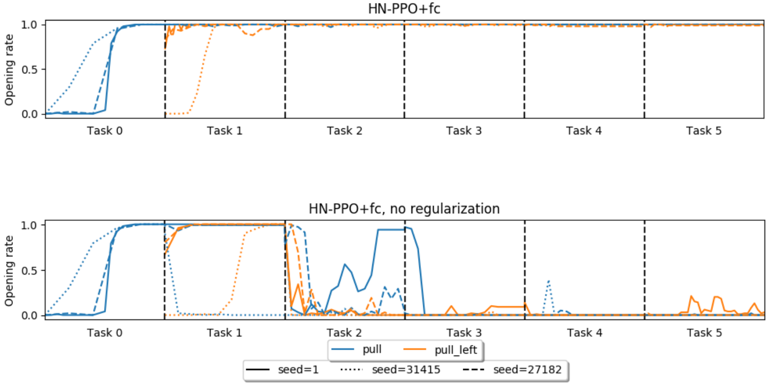
HN-PPO protects against catastrophic forgetting

Metric	PPO-finetuning	HN-PPO
Accuracy	0.20 ± 0.035	$\mathbf{0.81} \pm 0.041$
Remembering	0.47 ± 0.060	$\mathbf{1.00} \pm 0.0024$

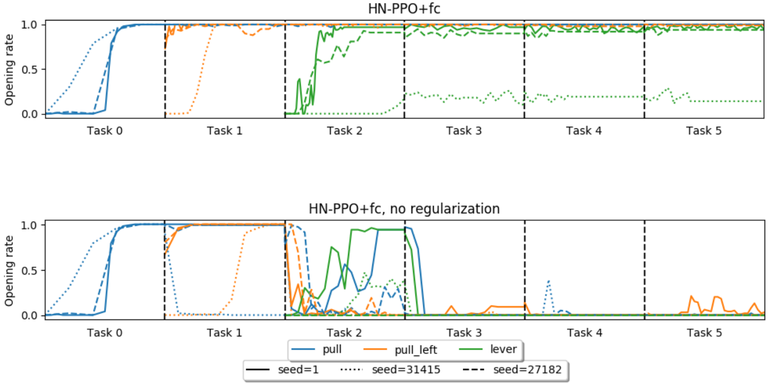
HN regularization is required for CL performance



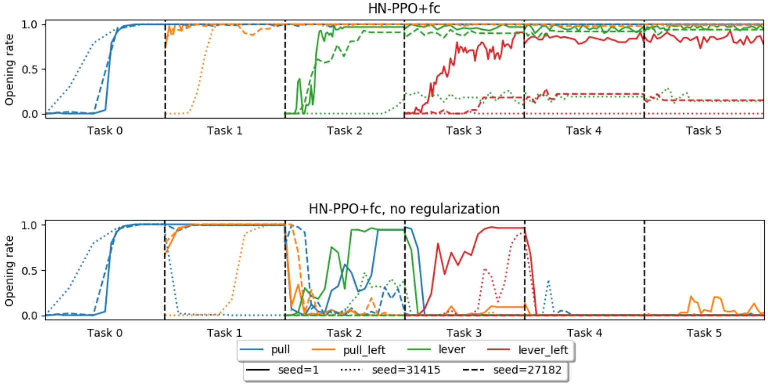
HN regularization is required for CL performance



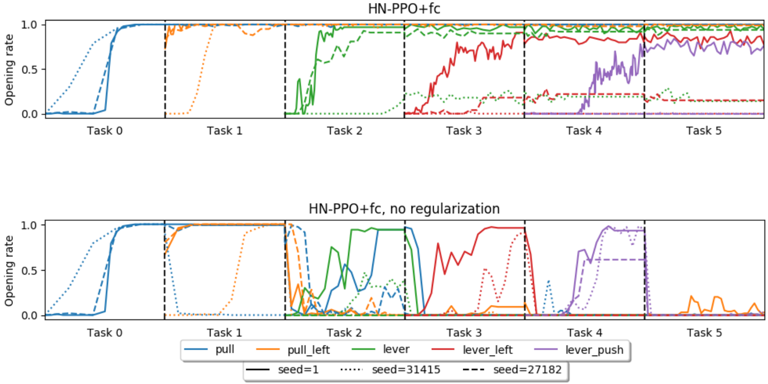
HN regularization is required for CL performance



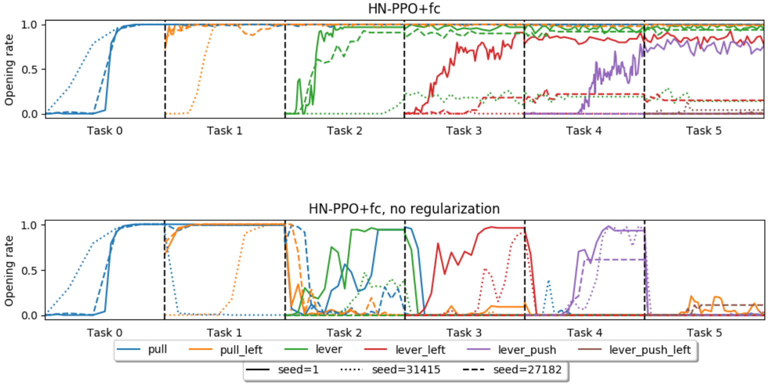
HN regularization is required for CL performance



HN regularization is required for CL performance



HN regularization is required for CL performance

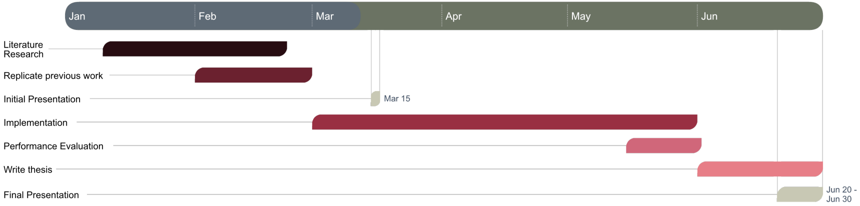


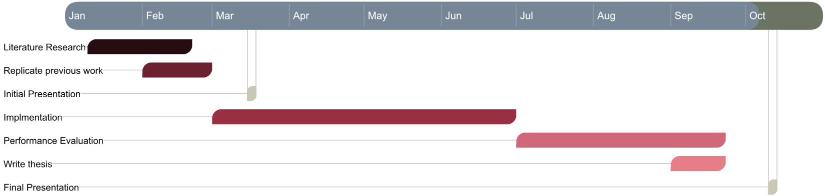
DoorGym demo

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
  - Previous task checkpoint dependence

Timeline


Jan
Feb
Mar
Apr
May
Jun
Literature Research
Replicate previous work
Initial Presentation
Mar 15
Implementation
Performance Evaluation
Write thesis
Final Presentation
Jun 20 - Jun 30


Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Literature Research
Replicate previous work
Initial Presentation
Implmentation
Performance Evaluation
Write thesis
Final Presentation