Hypernetwork-PPO for Continual Reinforcement Learning

Philemon Schöpf, Sayantan Auddy,
Jakob Hollenstein, Antonio Rodriguez-Sanchez

Intelligent and Interactive Systems Group
University of Innsbruck
Innsbruck, Austria

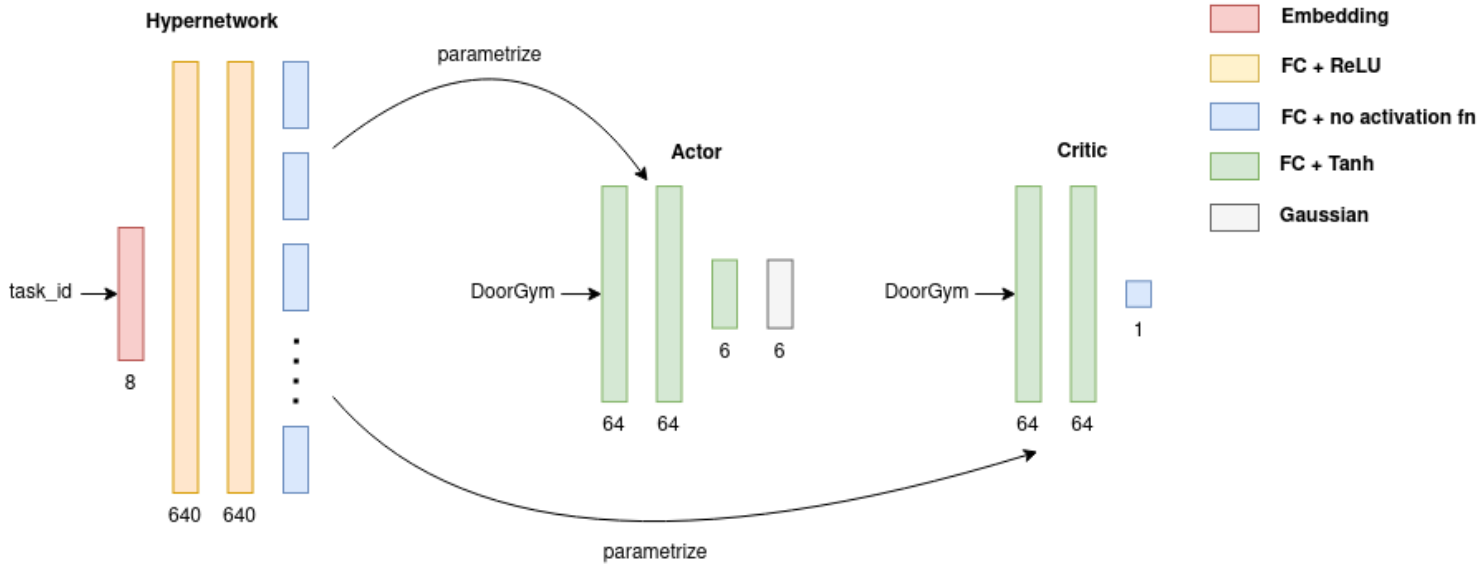
2022-11-10

Motivation

- Goal: Learn multiple tasks sequentially
- Cannot revisit old environment when learning new tasks
- Do not forget old skills

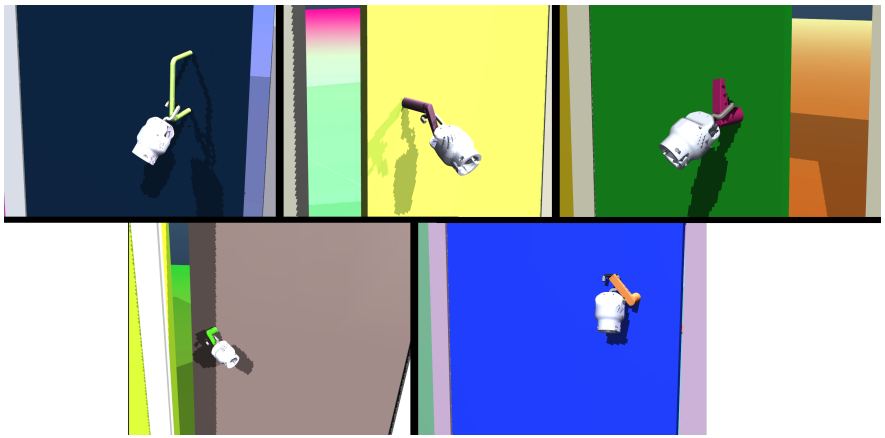
Hypernetworks

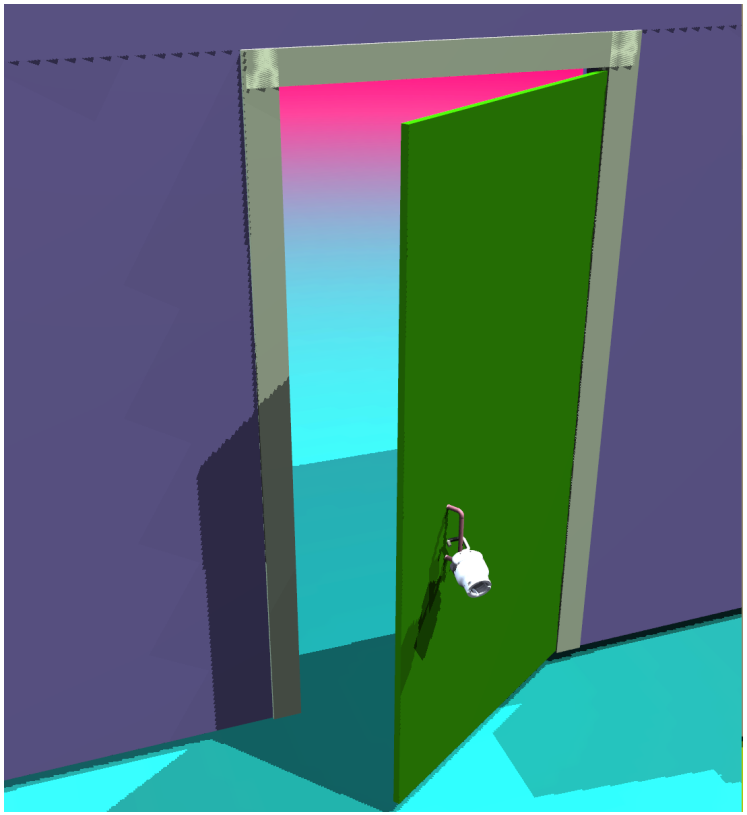
- Network that outputs a network
- Task ID as input
- Target networks determine policy/dynamics
- Regularization on changes of outputs for old tasks



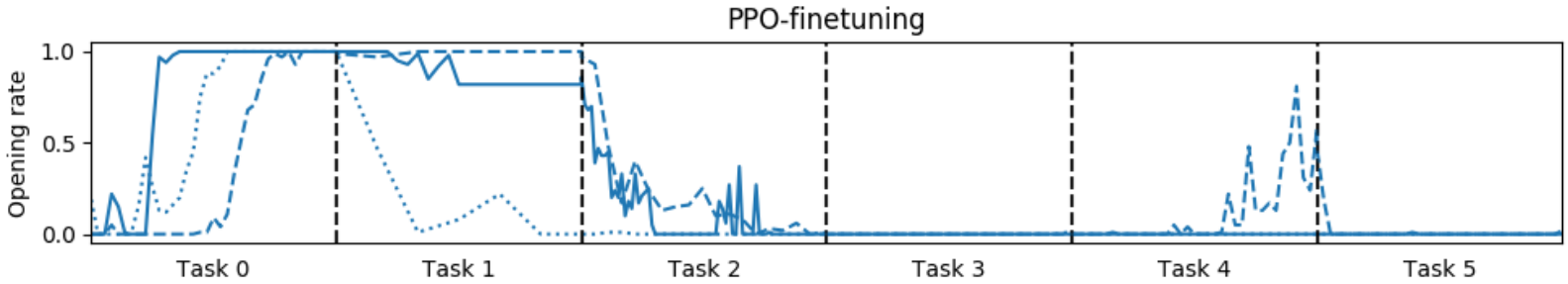
DoorGym

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

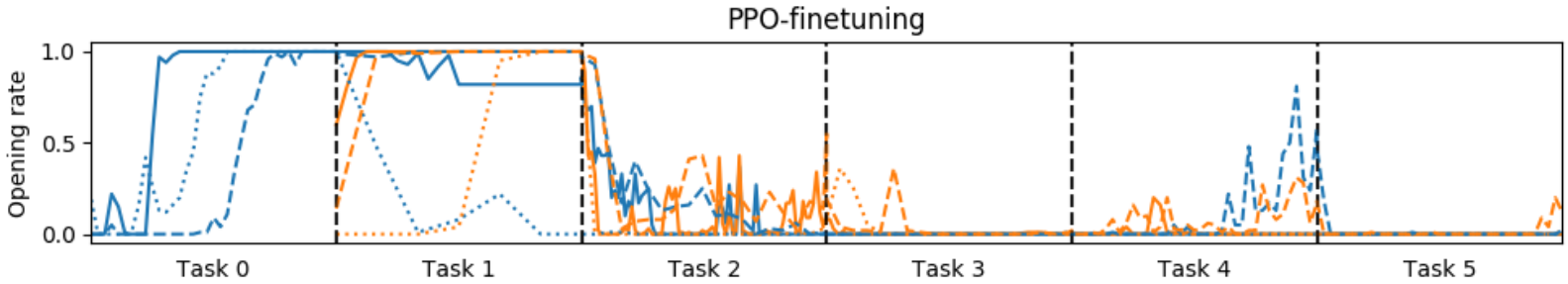




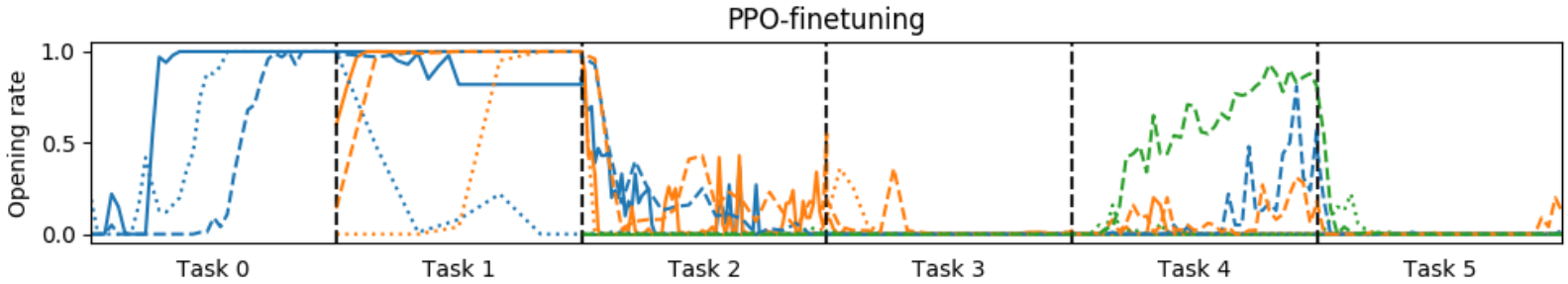
HN-PPO protects against catastrophic forgetting



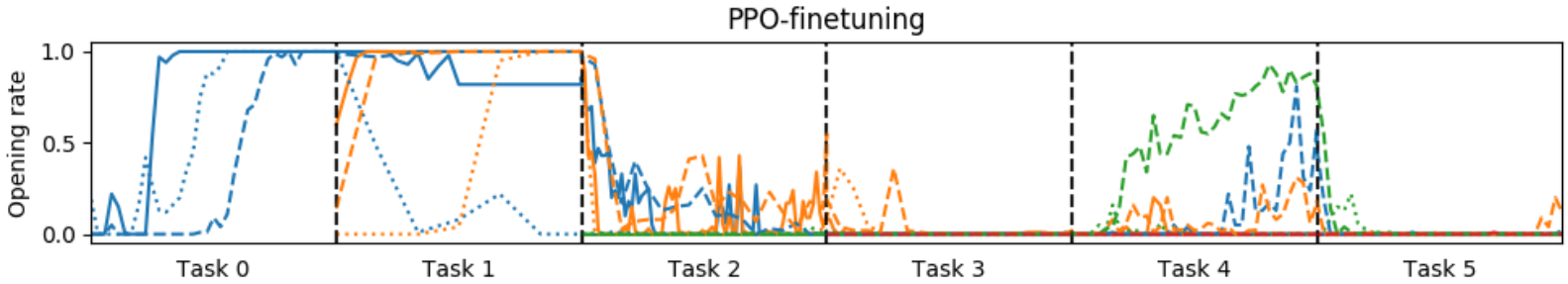
HN-PPO protects against catastrophic forgetting



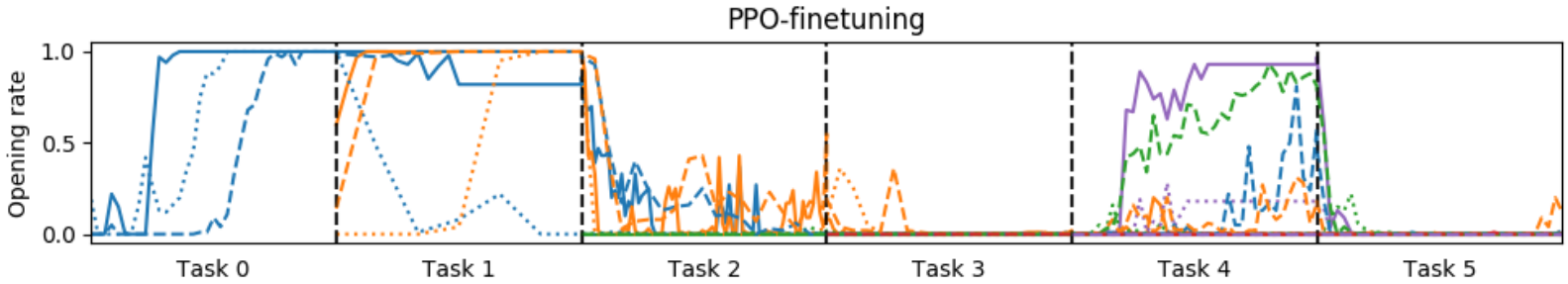
HN-PPO protects against catastrophic forgetting



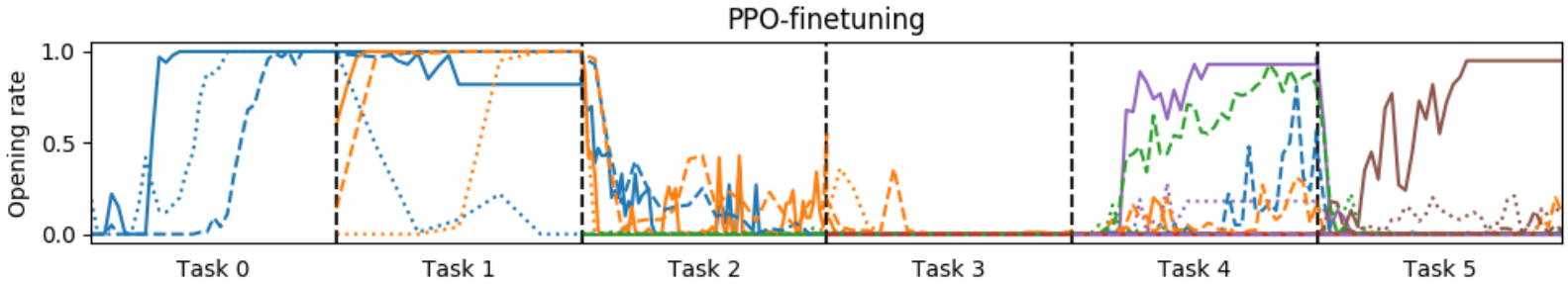
HN-PPO protects against catastrophic forgetting



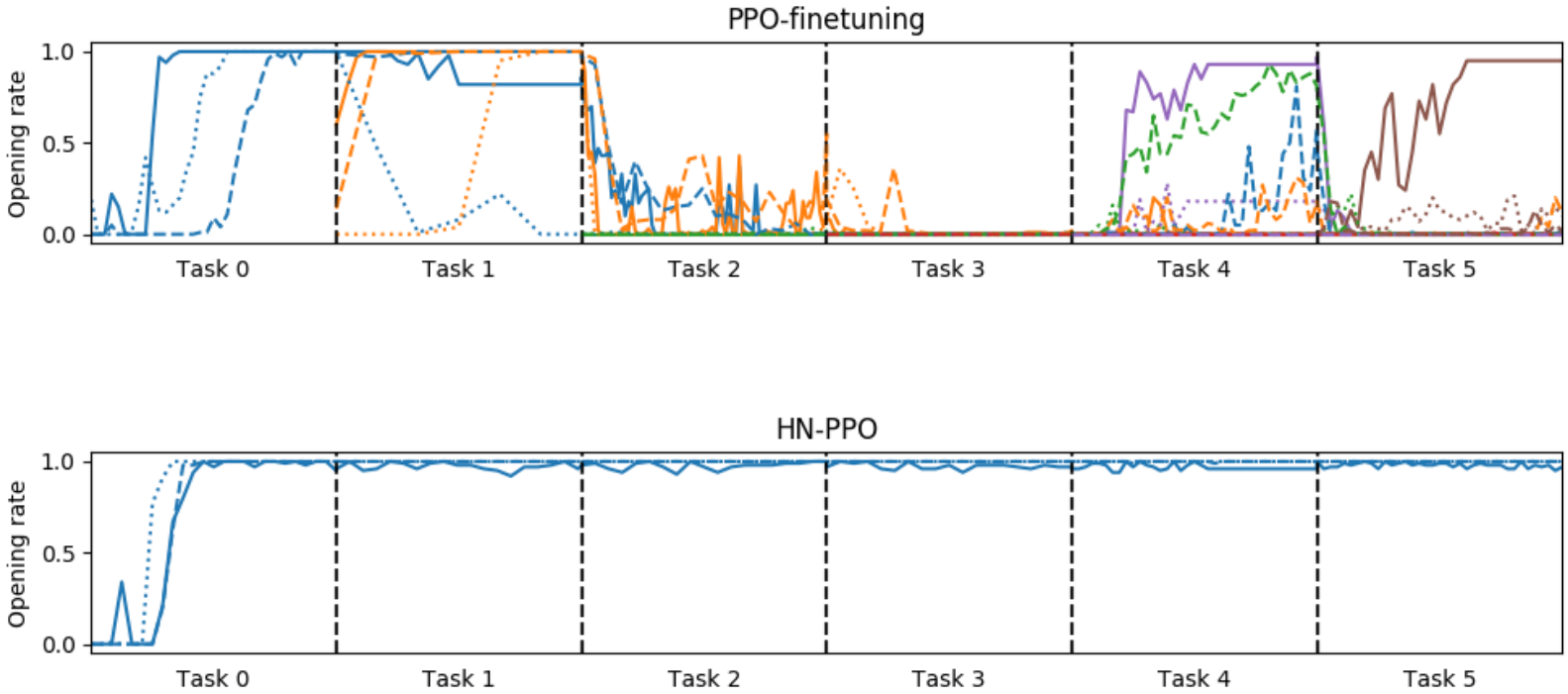
HN-PPO protects against catastrophic forgetting



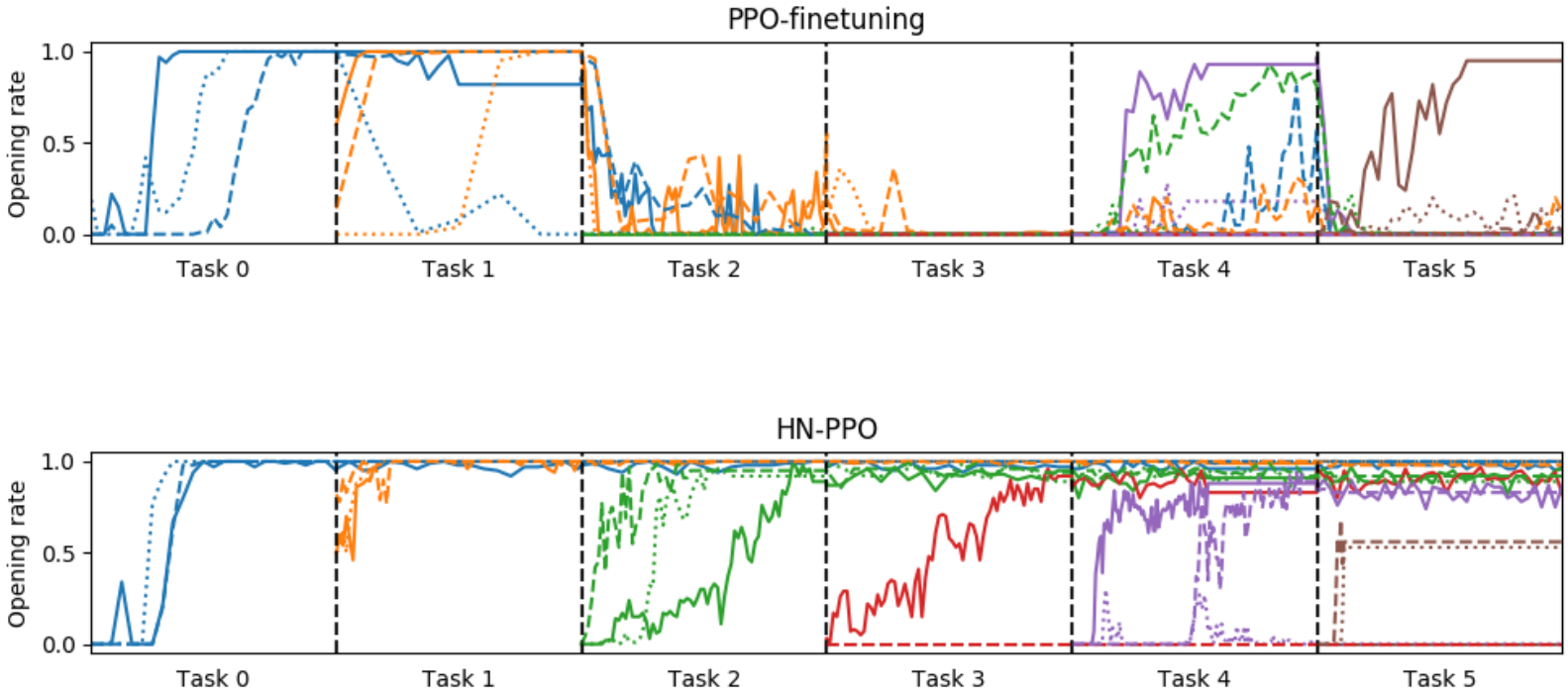
HN-PPO protects against catastrophic forgetting



HN-PPO protects against catastrophic forgetting



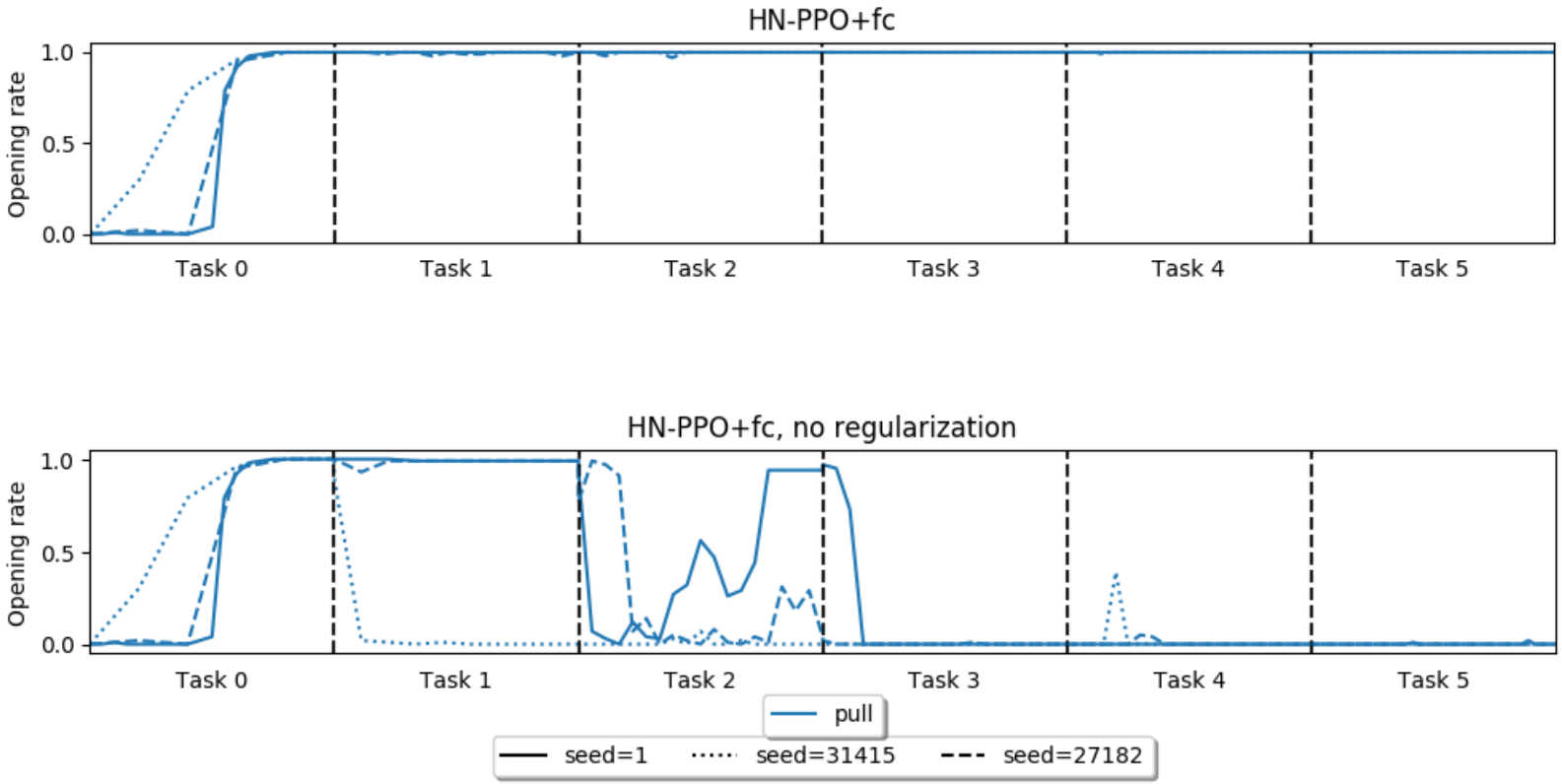
HN-PPO protects against catastrophic forgetting



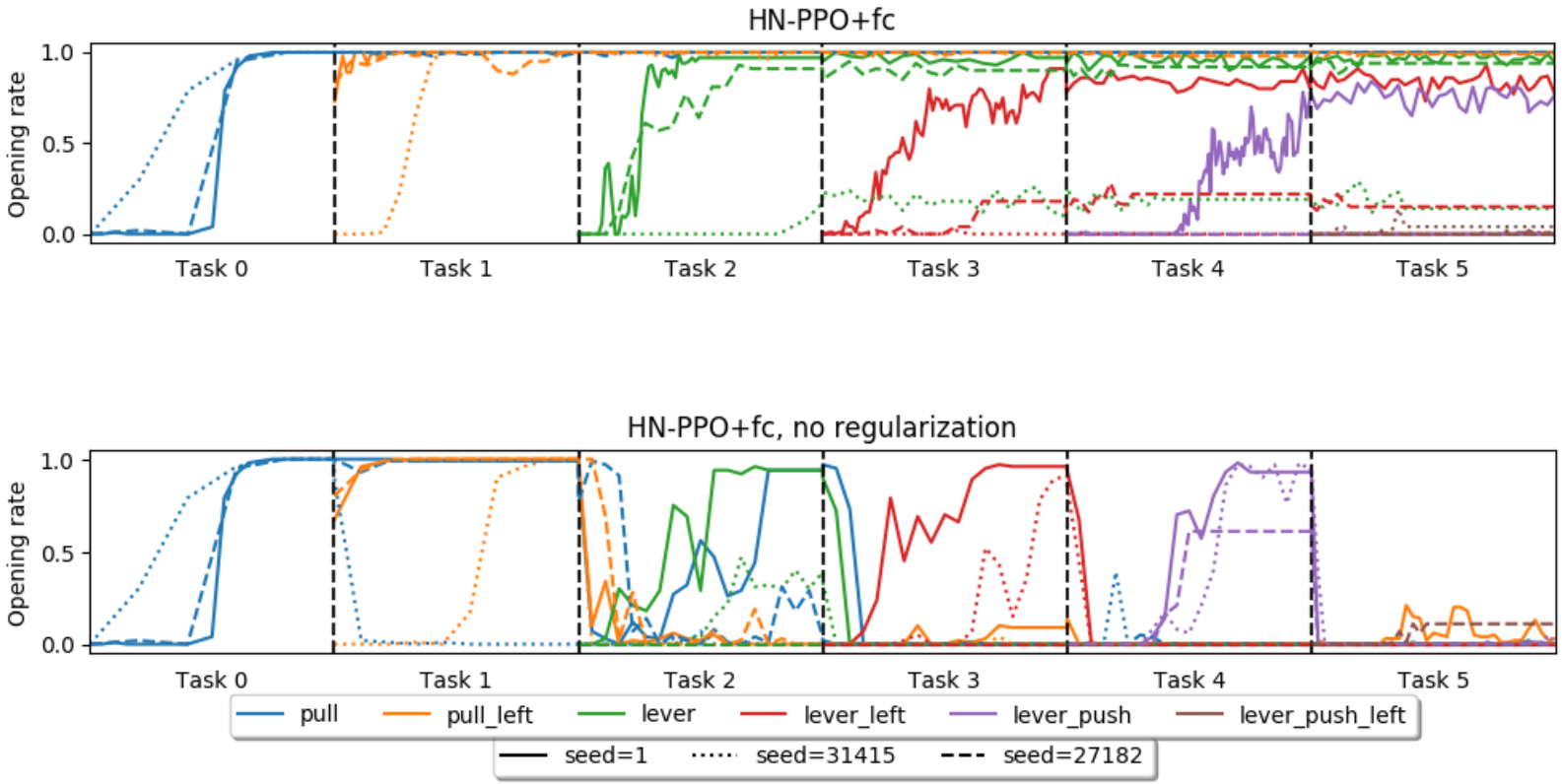
HN-PPO protects against catastrophic forgetting

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



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