Evaluating and Enhancing the Robustness of Proximal Policy Optimization to Test-Time Corruptions in Sequential Domains

Author:
Mate Rodić

Supervisor:
Mustafa Celikok

Responsible Professor:
Frans Oliehoek

I. Introduction

- Real world has noise, such as **observation noise** and action delays
- **Test-time corruptions** can induce drastic policy failures, yet are rarely evaluated in standard PPO benchmarks
- Prior work focuses on training stability in clean simulators, overlooking real-world uncertainties
- This paper quantifies **PPO's performance degradation under controlled noise** (σ up to 0.5) in **CartPole-v1 and Highway-env** environments
- This project evaluates **PPO's robustness** under such conditions and investigates whether simple techniques–like **using memory (LSTM)** or **training with noise**–can improve resilience.
- We conduct a comparison of baseline PPO to improved variations to conclude whether robustness can be improved, and what gives the best results

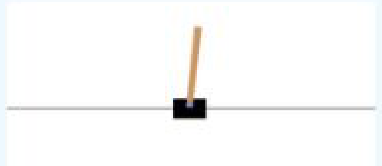

Figure 1. Cartpole environment

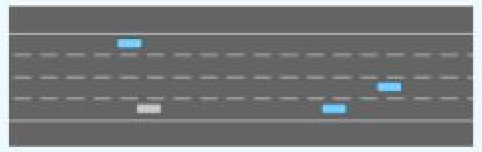

Figure 2. highway environment

II. Research Questions

1. **How does standard PPO performance degrade as test-time perturbations increase?**
2. **To what extent can recurrent architectures or noise-augmented training mitigate this degradation?**

III. Methodology

Algorithms

- Feed-forward PPO (baseline)
- Recurrent PPO (adds LSTM - long short-term memory, with size of 10)
- Noisy-PPO (Gaussian noise with σ= 0.1 during training)
- Recurrent-Noisy PPO (combines LSTM + noise injection)

Training setup

100k time steps, 5 seeds
2048 (4096 for recurrent) epochs
128 batch size
0.0003 learning rate
0.2 clip range

Hyperparameter Tuning

σ chosen to be 0.1 for noisy PPO

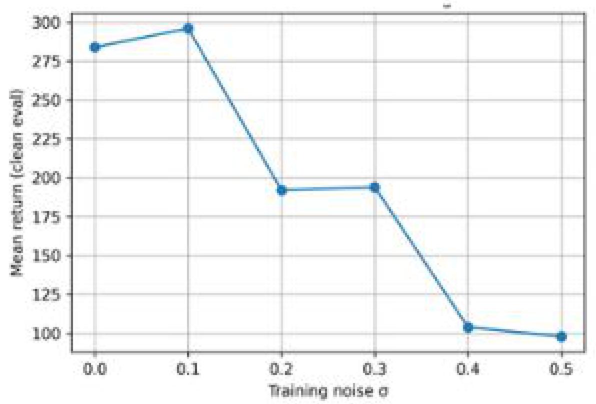

Figure 3. Mean return vs training noise

Test-time corruptions = Gaussian noise

σ ∈ {0.1, 0.2,0.3,0..4, 0.5})

Evaluation

- mean return
- standard deviation
- AUDC
- paired t-test to compare different to baseline model

IV. Results

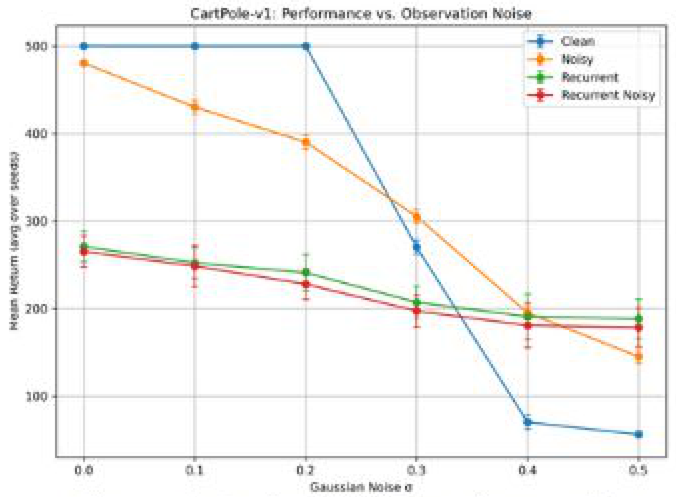

Figure 4. Performance vs observation noise - cartpole

In **cartpole PPO drops** 500→50 at σ=0.5, while **variants hold ~80%** of original (lower) performance

In **highway PPO collapses** after σ=0.2 and **other agents retain ~60%** at σ=0.5.

PPO **variations overall outperform baseline** PPO (higher AUDC), and the difference is more noticeable in a highway environment

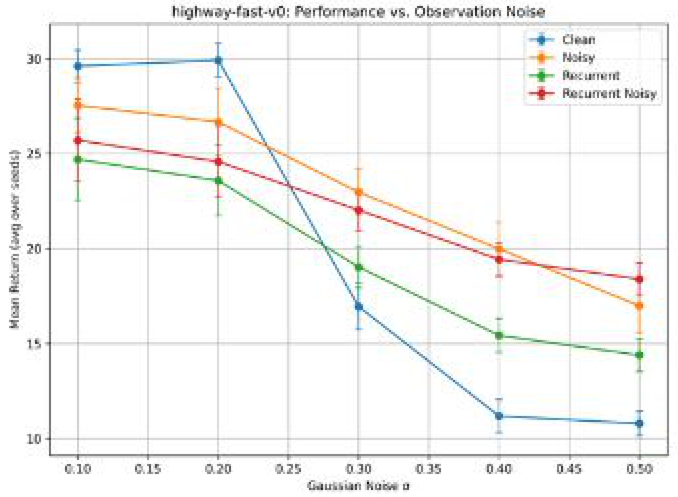

Figure 5. Performance vs observation noise - highway

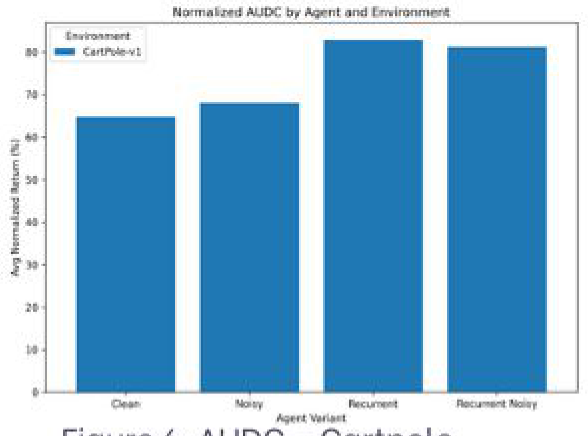

Figure 6. AUDC - Cartpole

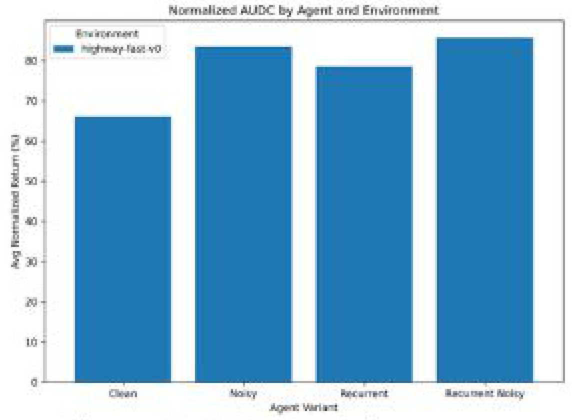

Figure 7. AUDC - Highway

V. Conclusion

- **Standard PPO agents are significantly affected** by small corruptions and performance drops are observed
- **LSTM memory (Recurrent PPO) adds more** stability when subjected to test-time noise, but experiences lower clean-environment performance
- **Noise-injection in training (Noisy-PPO) is a simple yet effective** way to reduce brittleness with minimal architecture changes
- **Combination of the two** (Recurrent Noisy-PPO) exhibits benefits of both variations and shows **slighly better performance**
- The **robustness has not been fully achieved**, as larger amounts of noise still affect all models.
- **Variations** of PPO exhibit **lower performance in clean test-time environment**, which could not have been mitigated

VI. Limitations & Future Work

- **Different environments** could be tested to see adaptability in other scenarios
- Only Gaussian observation noise was tested - **other real-world disturbances** like action delays can be used in training and testing
- **Training** for highway-env could have been **more extensive,** with **more timesteps** in order to get better models and possibly results

Contact Information: M.R.Rodic@student.tudelft.nl