Open RL Benchmark: Comprehensive Tracked Experiments for Reinforcement Learning

Have you ever seen the fancy learning curves from deep reinforcement learning (RL) papers? They look like this (copied from the PPO paper)

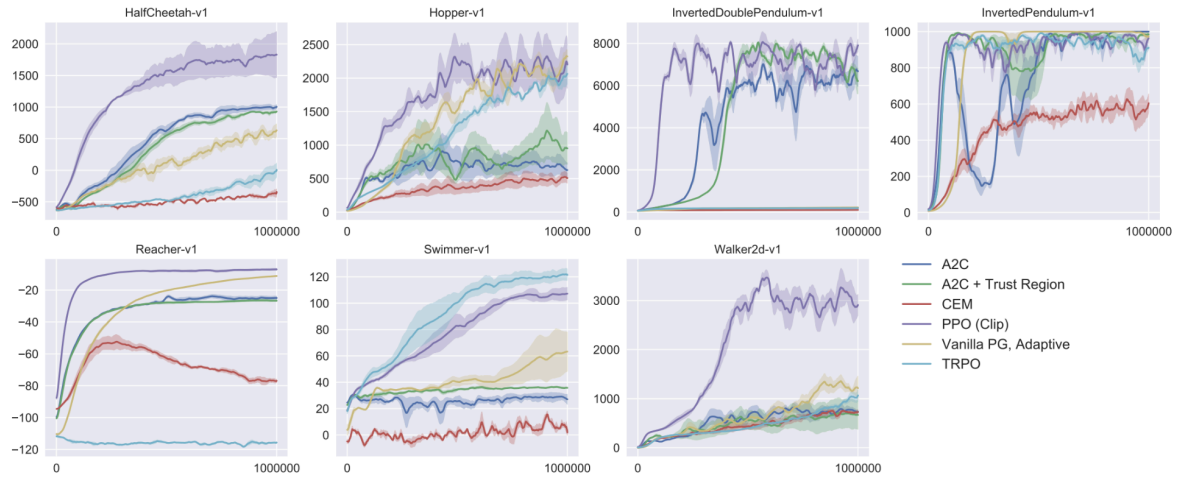


Figure 3: Comparison of several algorithms on several MuJoCo environments, training for one million timesteps.

They look really exciting for a minute, but how do you reproduce figures like this? Depending on the project, you may struggle

- finding relevant source code and hyperparameters
- finding reproduction instructions
- configuring environments with incomplete dependency versions
- comparing these figures with your figures by painfully playing with matplotlib…🤮
- visualizing what the agent is actually doing 🤖
- … the list goes on

This is a very common issue in deep RL and we need a more convenient resolution.

Open RL Benchmark

In this work, we propose the Open RL benchmark, a comprehensive collection of tracked deep RL experiments from popular RL libraries. Current participating libraries include SB3, Tianshou, CleanRL, but expect more to come. The following table contains the features comparison between data available in deep RL papers and Open RL Benchmark.

	Deep RL papers' figures or tables	Open RL Benchmark
Ability to "fork" metrics and compare them with users' own metrics	⛔	✅
Source code	🤔 maybe	✅
Pinned dependencies	🤔 maybe	✅
Complete list of hyper-parameters	✅ usually (however sometimes they spread over different pages)	✅
Exact command to reproduce results	🤔 maybe	✅
Training metrics	⛔ (usually just episodic return figure and no figure on losses or other metrics)	✅
System metrics	⛔ (rarely available)	✅
Customize x-axis for training metrics (such as using relative time or timesteps as the x-axis)	⛔	✅
Zoom in to the metrics	⛔	✅
Videos of agents' gameplay	🤔 maybe but usually no	🤔 maybe
Logs (stdout and stderr)	⛔	✅
Trained models	🤔 maybe but usually no	🤔 maybe (well… CleanRL doesn't have it, but Tianshou and SB3 should have it)
Downloadable metrics	⛔ (rarely available)	✅ (collaboration with https://github.com/wookayin/expt)

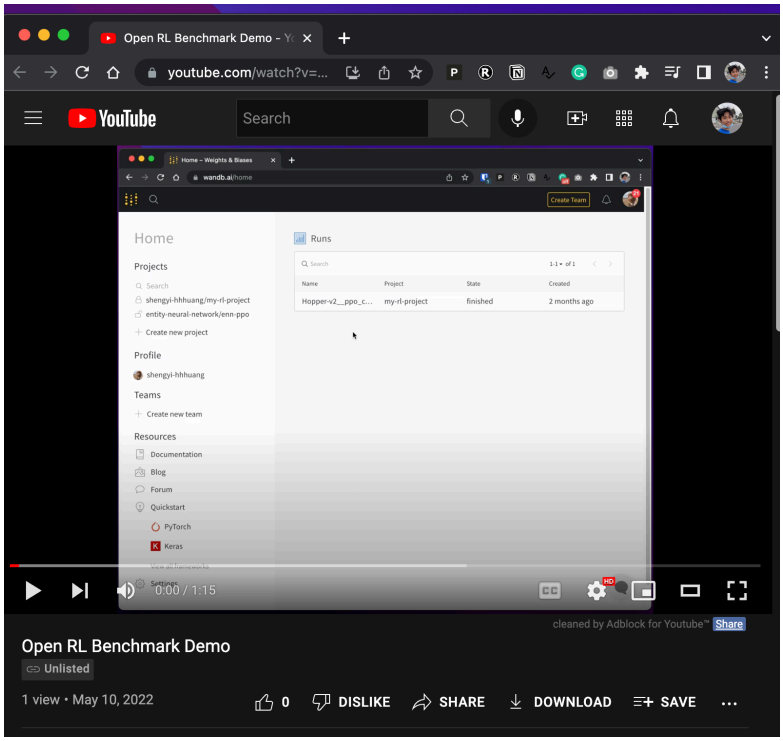
"Fork" Metrics: A Killer Feature

Open RL Benchmark has a killer feature that allows arbitrary users to compare their own metrics against those from ORLB. This is helpful when someone reproduces an existing algorithm and wants to compare performance against reputable implementations.

The following demo shows how to “fork metrics”, which has the following steps:

1. I implemented my PPO implementation named `ppo_continuous_action.py` and ran a tracked experiment in my W&B project named “my-rl-project”.
2. I opened Open RL Benchmark’s report that tracks openai/baselines’ PPO performance
3. I cloned this report in my W&B project named “my-rl-project”.
4. I added the “Run Set” from my own W&B project “my-rl-project” which **merged my tracked experiment with Open RL benchmark’s tracked experiments**
5. Finally, I was able to compare the performance of my PPO against openai/baselines’ PPO

See https://youtu.be/8PeBYUUJ5iw for a demo, where I cloned a report from someone else and compared my metrics with his.



Specification

All experiments from Open RL Benchmark follows this specification

1. Record all the hyper-parameters used for training, which must include
    - The name of the training environment
    - The name of the algorithm
2. Record all the raw unaveraged metrics in training or evaluation
3. If possible, Record the videos of the agents' gameplay throughout training (possibly via `gym.wrappers.RecordVideo` wrapper and `wandb.init(..., monitor_gym=True)`

4. Use wandb's tensorboard integration to do logging, therefore having a **common x-axis** named `global_step`
5. Have a PR that documents the reproduction scripts and instructions
6. Create a report that displays episodic return curves for experiments for each type of game
    - The report should link to the "reproduction instruction" code and PR