Benchmarking the Spectrum of Agent Capabilities

Danijar Hafner


1. Introducing Crafter

Crafter is an open-world survival game for RL research with and without rewards.

The player needs to collect resources and craft tools, all while struggling to survive.

The goal is to unlock 22 semantically meaningful achievements per episode.

Crafter evaluates a broad range of agent abilities within a single env and training run.

Pure Python, easy to install, use, and modify to maximize research productivity!

2. Research Challenges

- **Exploration** Unlocking many achievements in the technology tree requires both deep and wide exploration.
- **Generalization** Every episode uses a different procedurally generated world, so agents need to detect similar situations.
- **Reusability** Skills need to repeatedly collect water, forage for food, and collect basic resources such as wood and stone.
- **Credit assignment** Several achievements feature long temporal dependencies, such as waiting for a plant to grow fruits to eat.
- **Memory** Agent needs to remember lakes to repeatedly find water and remember where it has already been to find rare resources.
- **Representation** Agent-centric visual observations that change with the day-night cycle require learning stable representations.
- **Survival** Having to find water, food, sleep, and defend against monsters helps prevent trivial solutions of unsupervised agents.

3. Achievements

The goal of Crafter is to unlock **22 achievements per episode**, which correspond to meaningful milestones in agent behavior and measure a diverse range of abilities of agents with or without reward.

4. Agent Ability Spectrum

Success rates are computed per achievement across all training episodes leading up to the budget of 1M env steps. This offers insights into an agent’s strengths and weaknesses.

5. Benchmark Scores

Crafter is of **appropriate difficulty**, allowing current top agents to make some learning progress while posing a substantial research challenge to reach **human performance** in the future.

The benchmark score is the **geometric mean** over success rates:

This is useful when tasks are of different difficulty or reward scale, without having to know or assume the difficulties a priori. Geometric mean (mean in log-space) weighs difficult tasks stronger, rewarding agents for demonstrating broad abilities over many tasks. Collecting diamonds 1% of the episodes instead of 0% is a big improvement. Collecting wood 95% instead of 90% is not.

6. Available Resources

Play Crafter Yourself!

Supports Linux, Mac, Windows

- `python3 -m pip install crafter`
- `python3 -m pip install pygame`
- `python3 -m crafter.run_gui`

Useful resources are available on the project website:

- **Human expert video**
- **Emergent agent behaviors**
- **Baseline implementations (Docker)**
- **Baseline scores (JSON)**
- **Plotting scripts**
- **Human dataset (NPZ)**

Website with videos, data, and code: danijar.com/crafter