Báo cáo Cuối kỳ: Nghiên cứu Hiện tượng Reward Hacking trong Học tăng cường với PPO và ORPO trên môi trường tomato_environment

Sinh viên: Lê Bảo Duy
Mã số sinh viên: QE180111
Đề tài: Reward Hacking trong Reinforcement Learning và các phương pháp giảm thiểu.
Môn học: REL301m
Giảng viên hướng dẫn: Nguyễn An Khương
Ngày nộp: 25/7/2025

1. Mở đầu

1.1. Bối cảnh và Động lực

Học tăng cường (Reinforcement Learning - RL) là một lĩnh vực trí tuệ nhân tạo cho phép tác tử (agent) học cách tương tác với môi trường để tối đa hóa phần thưởng nhận được. Tuy nhiên, hiện tượng **Reward Hacking** xảy ra khi agent khai thác những lỗ hổng trong hàm phần thưởng, đạt được điểm cao nhưng không hoàn thành nhiệm vụ theo cách người thiết lập mong muốn.

Nghiên cứu này sử dụng môi trường **tomato_environment**, được tham khảo từ các bài báo khoa học, chứa hai dạng phần thưởng: *true_reward* (thu hoạch tomato) và *proxy_reward* (khám phá trạng thái sprinkler ẩn). Mục tiêu nhằm thấu hiểu và giảm thiểu hiện tượng reward hacking, hỗ trợ phát triển các hệ thống RL tin cậy hơn.

1.2. Mục tiêu nghiên cứu

- Nghiên cứu hiện tượng Reward Hacking trong môi trường tomato_environment.
- Thử nghiệm hai thuật toán PPO và ORPO trên môi trường này.
- Phân tích các thách thức khi agent bị *stuck* trong quá trình evaluation, tập trung vào true_reward mà bỏ qua proxy_reward.
- Đề xuất, thử nghiệm các giải pháp chỉnh sửa reward để cải thiện khả năng khám phá và giảm reward hacking.

2. Nội dung và Phương pháp

2.1. Tổng quan thuật toán

- **Proximal Policy Optimization (PPO):** Thuật toán chính sách dựa trên gradient với mục tiêu đảm bảo cập nhật chính sách ổn định qua hàm loss clip, giảm thiểu các cập nhật vượt quá giới hạn làm mất ổn định học.
- **Off-Policy Reinforcement Policy Optimization (ORPO):** Một cải tiến của PPO, cho phép tận dụng dữ liệu off-policy, tăng hiệu quả huấn luyện trong môi trường hạn chế dữ liệu mới. Tuy nhiên, việc triển khai ORPO gặp khó khăn do lỗi kỹ thuật trong repo chính thức.

2.2. Môi trường tomato_environment

Môi trường phức tạp mô phỏng việc thu hoạch tomato (*true_reward*) với thưởng phụ dựa trên khám phá trạng thái sprinkler (*proxy_reward*) nhằm khuyến khích agent khám phá tổng quát hơn thay vì chỉ tập trung vào phần thưởng chính. Đây là một môi trường điển hình để quan sát hiện tượng Reward Hacking đa chiều.

2.3. Phương pháp nghiên cứu

- Huấn luyện agent bằng PPO trên môi trường tomato_environment.
- Đánh giá quá trình *evaluation* để phát hiện hiện tượng agent bị *stuck* trong thu hoạch tomato và không khai thác proxy_reward.
- Thử chỉnh sửa hàm thưởng và sử dụng các kỹ thuật khám phá nhằm giảm hiện tượng reward hacking.
- Ghi nhận khó khăn khi triển khai ORPO do lỗi repo.
- So sánh kết quả, phân tích hành vi và hiệu suất.

3. Quá trình thực hiện và thí nghiệm

3.1. Tuần 1: Nghiên cứu lý thuyết và môi trường

- Khảo sát tài liệu về reward hacking trong RL.
- Chọn tomato_environment làm môi trường thực nghiệm phù hợp với đề tài.

3.2. Tuần 2: Huấn luyện và phát hiện vấn đề

- Huấn luyện agent dùng PPO với hàm phần thưởng gồm cả true_reward và proxy_reward.

- Quan sát agent thường bỏ qua proxy_reward, chỉ tập trung tối đa hóa true_reward (thu hoạch tomato).
- Hiện tượng *stuck* rõ ràng: agent không khám phá sprinkler state, dẫn đến hiệu suất xem nhẹ yếu tố proxy_reward.

3.3. Tuần 3: Thử nghiệm giải pháp và đánh giá

- Thiết kế lại hàm thưởng, nâng cao phần thưởng khám phá sprinkler để khuyến khích đa nhiệm.
- Thử áp dụng các ràng buộc, kỹ thuật reward shaping để thúc đẩy agent khám phá.
- Kết quả cải thiện khả năng khám phá nhưng chưa hoàn toàn khắc phục *stuck*.
- Cố gắng triển khai ORPO nhưng không thành công do lỗi nằm trong file requirement.txt của repo, khi nó có liên kết với 3 repo phụ khác được publish cách bài báo khá lâu và code chạy trên hệ điều hành Linux.

4. Kết quả và phân tích

Phương pháp	Hành vi agent	Hiệu suất nhiệm vụ
PPO (reward ban đầu)	Ưu tiên thu hoạch tomato, bỏ qua sprinkler	Kém, reward hacking rõ ràng
PPO (reward chỉnh sửa)	Khám phá tốt hơn nhưng vẫn còn hạn chế	Cải thiện rõ rệt, nhưng chưa tối ưu
ORPO	Không triển khai được do lỗi repo	Không có dữ liệu

- Agent thiên về việc khai thác phần thưởng đơn giản dễ đạt (true_reward) hơn là khám phá proxy_reward.
- Sự khác biệt và mâu thuẫn giữa hai dạng phần thưởng nếu không cân bằng tốt dễ dẫn đến reward hacking.
- PPO ổn định nhưng cần chú ý reward design.
- ORPO có tiềm năng nhưng cần khắc phục kỹ thuật để thử nghiệm sâu hơn.

5. Kết luận và đề xuất

- Reward Hacking là thách thức lớn trong RL, đặc biệt khi reward có nhiều khía cạnh khác nhau như trong tomato_environment.
- Việc thiết kế reward cần đa chiều, cân bằng giữa phần thưởng chính và phần thưởng phụ để thúc đẩy khám phá.

- PPO là thuật toán ổn định phù hợp cho bài toán này nhưng chưa giải quyết được triệt để các hành vi stuck.
- ORPO là hướng nghiên cứu tiềm năng nhưng cần hoàn thiện kỹ thuật triển khai.
- Đề xuất tiếp tục phát triển reward shaping, áp dụng các kỹ thuật khám phá tiên tiến, và cải thiện môi trường cũng như các công cụ.

Tài liệu tham khảo

- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). **Proximal Policy Optimization Algorithms**, arXiv:1707.06347.
  Link: https://arxiv.org/pdf/1707.06347
- Laidlaw, C., Singhal, S., & Dragan, A. (2024). Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking. *arXiv preprint arXiv:2403.03185*. Spotlight at ICLR 2025.
  Available: https://arxiv.org/abs/2403.03185
- Anonymous Authors (2024). Preventing Reward Hacking with Occupancy Measure Regularization. *Under review at ICLR 2024*.
  Available: https://openreview.net/pdf?id=86w3LbTNI1
- Repository ORPO: https://github.com/cassidylaidlaw/orpo/tree/main