Reinforcement Learning final assignment: knights, archers and zombies

José Manuel Ros Rodrigo

Department of Advanced Computing Sciences
Maastricht University

08/06/2024

Summary

- PettingZoo's Knights Archers Zombies [TBG+21].
- Multi-agent environment.
- Proximal Policy Optimization (PPO).
- Experiment tracking with Weights & Biases [Bie20].

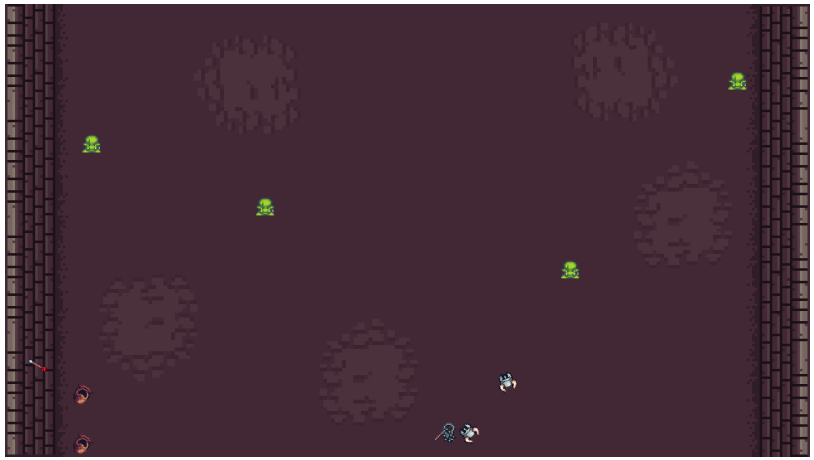

Table of contents

KAZ: rules and simplifications

- Game finishes when no agents remain alive or when a zombie reaches the bottom row.
- 6 different actions per agent.
- reward of $+1$ for killing a zombie.
- To simplify, one archer and one knight.
- To simplify, agents cannot be killed.
- Observation space $\rightarrow 24 \times 5$ matrix.
- max_zombies$= 10$, spawn_rate$= 20$, max_arrows$= 10$, max_cycles$= 900$.

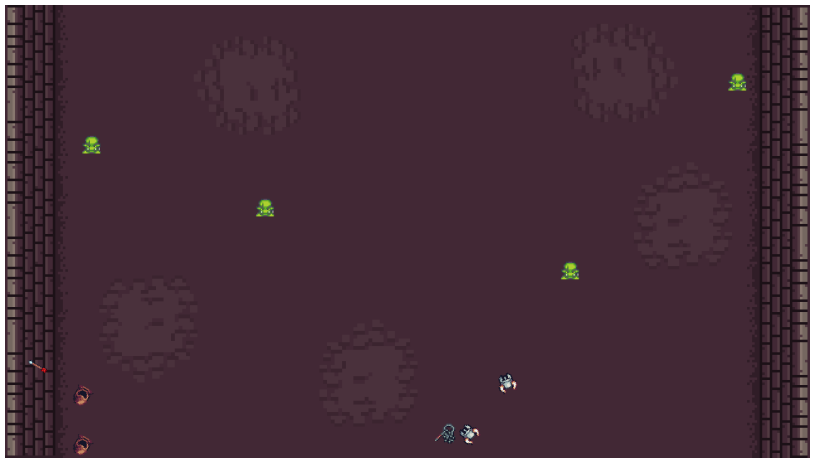

Table of contents

PPO implementation

- Based on CleanRL's implementation [HDY+22].
- All implementation details used but value loss clipping (9^{th}) [HDR+22].
- Two 1-hidden layer neural networks with 256 units per agent.
- Freeze and play (pray) scheme: started with random knight and train archer, then train knight and freeze archer, then loop.
- Needed to adapt it to the environment.

PPO: adaptation to KAZ

- KAZ is a multi-agent environment that doesn't have an API for vectorization.
- Created a class called *Envs* as a wrapper for the environments.
- Used SuperSuit's *SyncAECVectorEnv* class to have an Agent-Environment-Cycle that facilitates full control over the environments [TBH20].

Table of contents

Net size

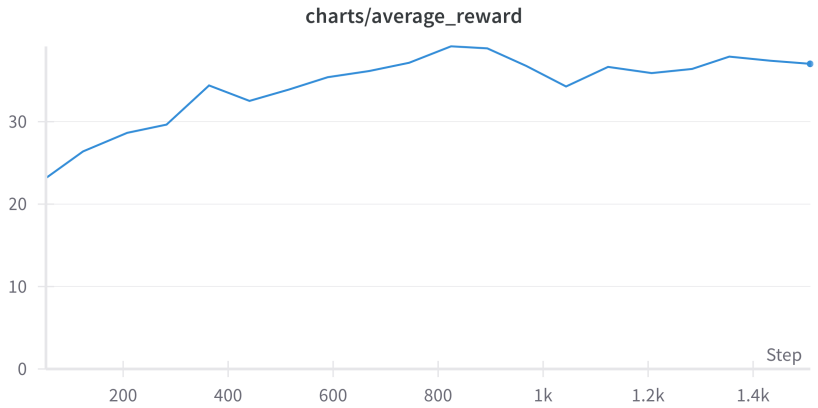

Figure: Small network (64 units).

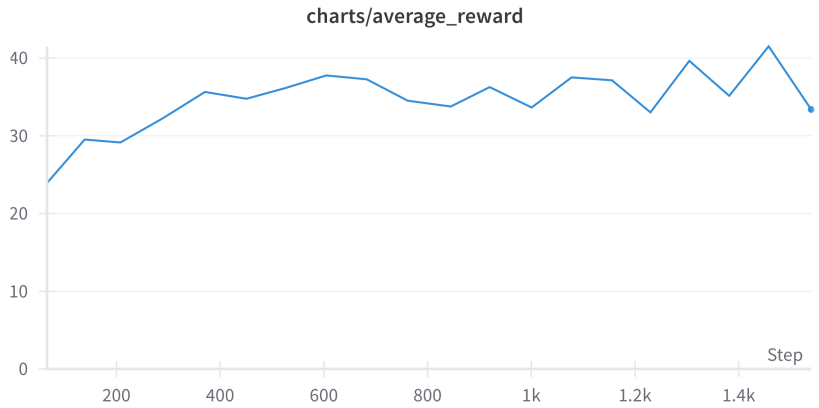

Figure: Big network (256 units).

Hyperparameters

- Goal $\rightarrow$ set with a good balance between training times and performance.
- Played with them in the first W&B runs (and in some crashed and deleted previous runs).
- Annealed learning rate from 4×10^{-4}, batch size of 16384, 8 environments, 327.680 total timesteps, entropy allowed, gamma of 1, rest of parameters by default.
- Learning rate of 10^{-3} ended up causing divergence.
- Approx. 1 hour 20 minutes per agent-training cycle.

Table of contents

Average rewards

Figure: Average rewards for the archer.

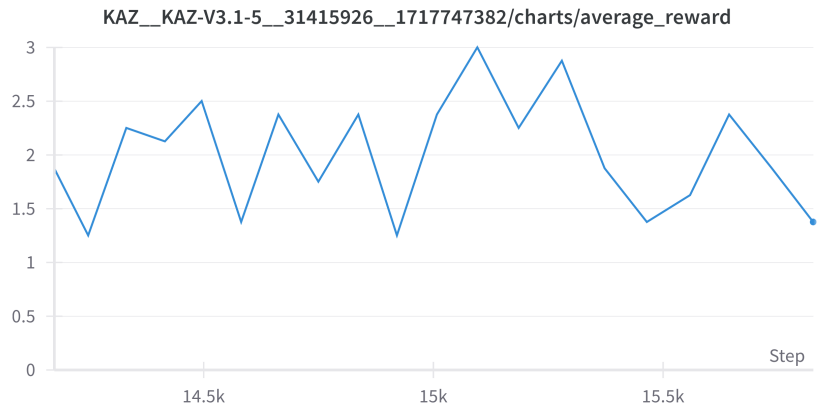

Figure: Average rewards for the knight.

Rewards are from the last cycles of the training loop. On average, this rewards are from $6 - 10$ different runs on the same environment.

Policy in action

W&B workspace

Table of contents

Links to the code & experiments

- Code on my GitHub.
- https://github.com/joros244/KAZ
- Experiments (and models) on Weights & Biases.
- https://wandb.ai/rl2024-umdacs/KAZ-RL?nw=nwuserjoros

Thank you for your attention!

References I

- Lukas Biewald, *Experiment tracking with weights and biases*, 2020, Software available from wandb.com.
- Kurt Driessens and Dennis Soemers, *Reinforcement learning course's slides*, 2024.
- Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang, *The 37 implementation details of proximal policy optimization*, ICLR Blog Track, 2022, https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.
- Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo, *Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms*, Journal of Machine Learning Research **23** (2022), no. 274, 1–18.
- Oleg Latypov, *A comprehensive guide to proximal policy optimization (ppo) in ai*, 2023, https://medium.com/@oleglatypov/a-comprehensive-guide-to-proximal-policy-optimization-ppo-in-ai-82edab5db200.

References II

- Brian Pulfer, *Ppo — intuitive guide to state-of-the-art reinforcement learning*, 2022, https://medium.com/@brianpulfer/ppo-intuitive-guide-to-state-of-the-art-reinforcement-learning-410a41cb675b.
- Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann, *Stable-baselines3: Reliable reinforcement learning implementations*, Journal of Machine Learning Research **22** (2021), no. 268, 1–8.
- J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al., *Pettingzoo: Gym for multi-agent reinforcement learning*, Advances in Neural Information Processing Systems **34** (2021), 15032–15043.
- J. K Terry, Benjamin Black, and Ananth Hari, *Supersuit: Simple microwrappers for reinforcement learning environments*, arXiv preprint arXiv:2008.08932 (2020).

References III

 Eric Yang Yu, *Coding ppo from scratch with pytorch*, 2020, https://medium.com/analytics-vidhya/coding-ppo-from-scratch-with-pytorch-part-1-4-613dfc1b14c8.