Exploration of sparse and dense reward structures in Google Research Football

Lakshay Dahiya, Steven Grisafi, Melvin Moriniere

University at Buffalo The State University of New York

ENVIRONMENT

Environment

- Stochastic environment developed by Google Research
- Complete football (soccer) simulation built for RL
- Different **rewards structures**
  - Sparse : 0 or 1 when scoring
  - Dense: 0 to 1 depending on distance to the goal and +1 when scoring
  - -1 reward if opposing team gets the ball
- Discrete **observations space** (115 dimension vector or pixel array)
- **19 discrete actions**

Environment scenario examples

- **academy_empty_goal** - Our player starts in the middle of the field with the ball, and needs to score against an empty goal.
- **academy_run_to_score** - Our player starts in the middle of the field with the ball, and needs to score against an empty goal. Five opponent players chase ours from behind.
- **academy_pass_and_shoot_with_keeper** - Two of our players try to score from the edge of the box, one is on the side with the ball, and next to a defender. The other is at the center, unmarked, and facing the opponent keeper.
- **academy_single_goal_versus_lazy** - Full 11 versus 11 games, where the opponents cannot move but they can only intercept the ball if it is close enough to them. Our center back defender has the ball at first.

What did we aim to solve?

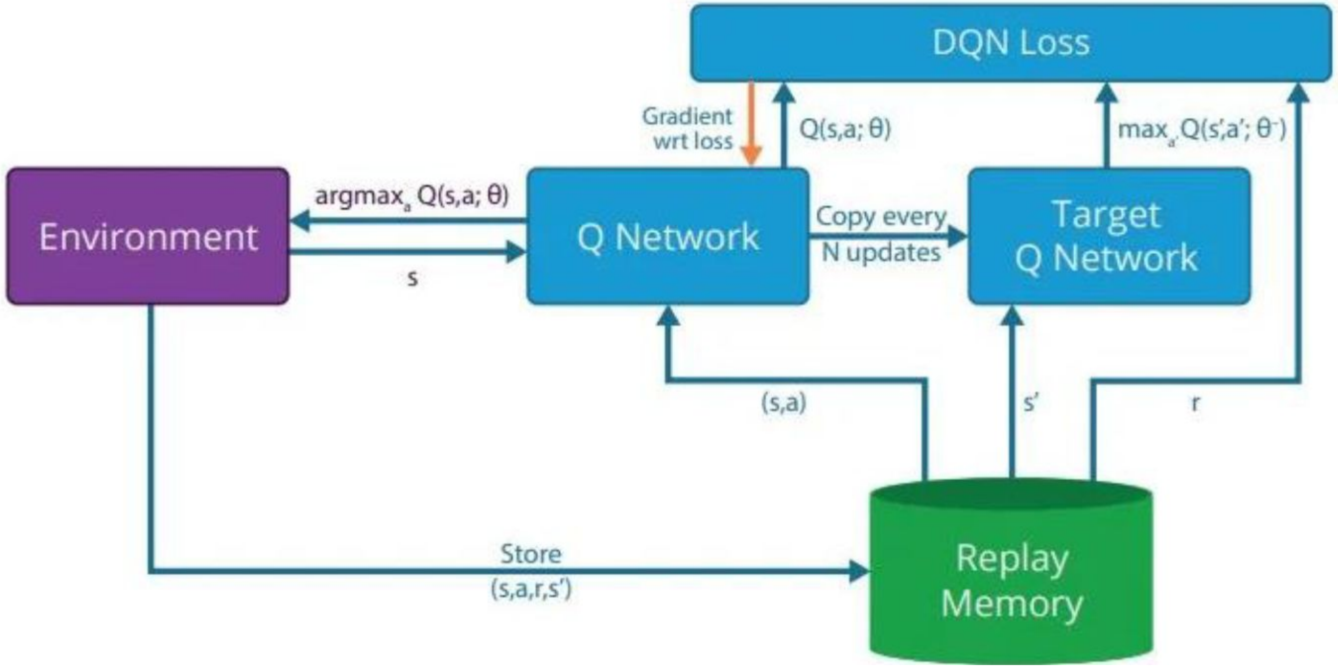
Solved environments with multiple algorithms and reward structures

- Applied algorithms to solve environments in the Football Academy sequence of progressively more complex environments.
- Applied DDQN to successfully solve 3 environments: empty goal (close), empty goal, run to score.
- DDQN was unable to solve pass and shoot with keeper.
- Used PPO to solve the most complex environment in Football Academy (single goal vs lazy)
- Compared results of sparse PPO training (reward at end of goal attempt) to dense PPO training (also adds smaller rewards for moving the ball down the field).

ALGORITHMS

DDQN

- 2 networks:
  - Target network and Q network → **reduce the variance**
- 1 replay Memory
  - store the episodes → **break correlation**
- Clip rewards
  - Ensure **gradients are well conditioned**

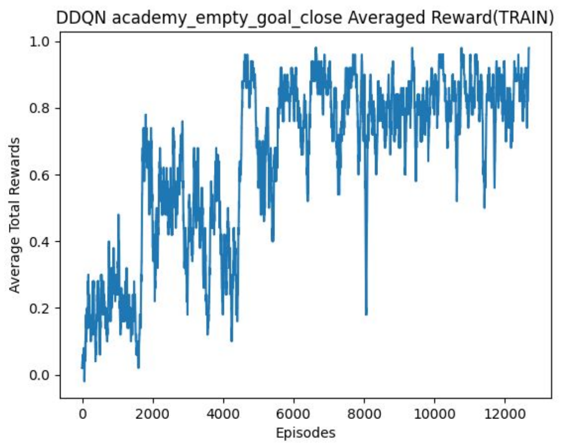


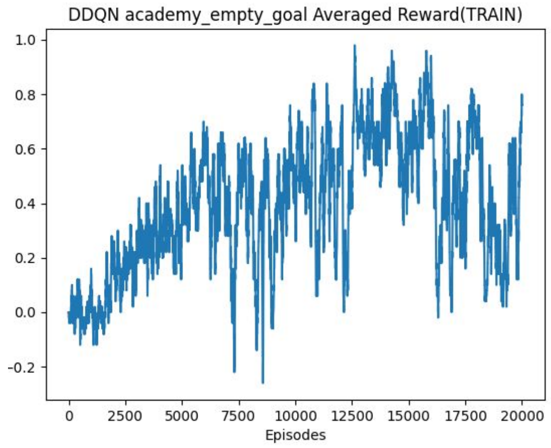
PPO

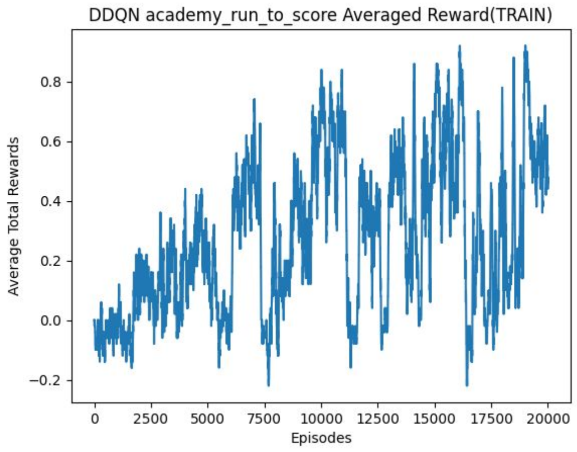
- DDQN began to take an infeasible amount of training time (approx. 3 days, 100k episodes without solving moderately complex environments).
- At this point, we chose to switch to an actor-critic method.
- PPO is an actor-critic method with the ratio of log policies between current and previous states clipped by an defined value.

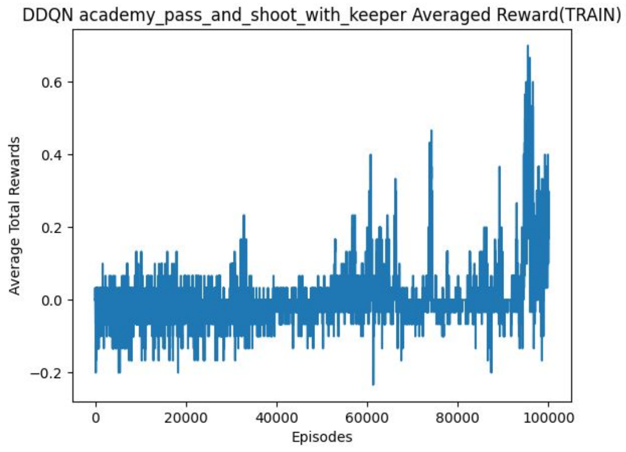
RESULTS

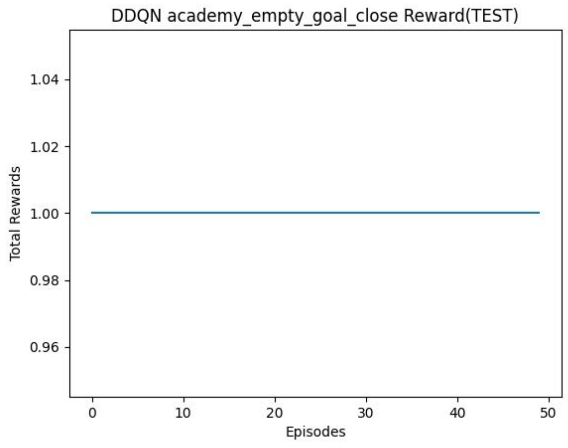
DDQN - Sparse Rewards

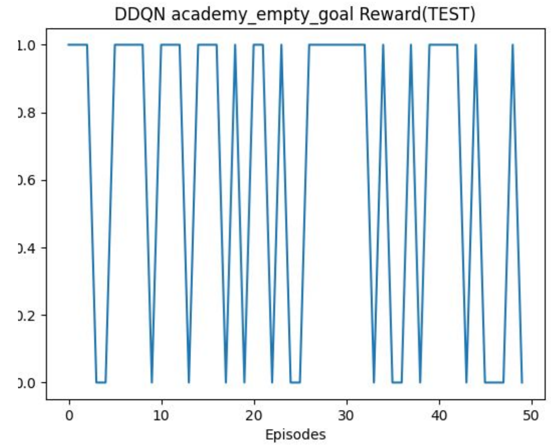


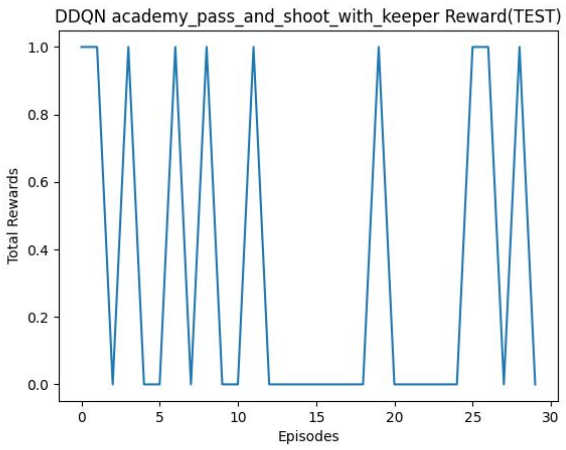




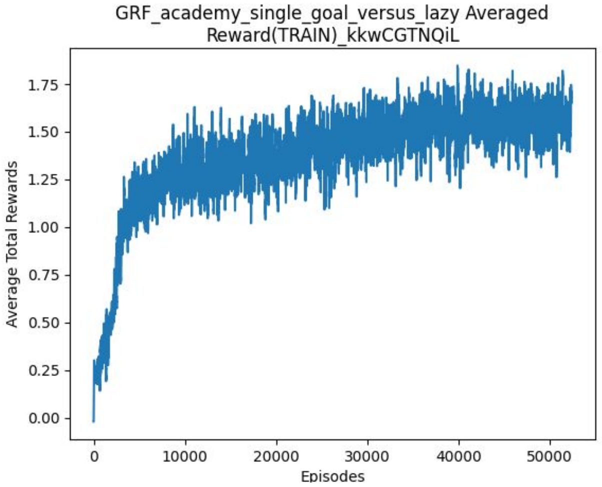


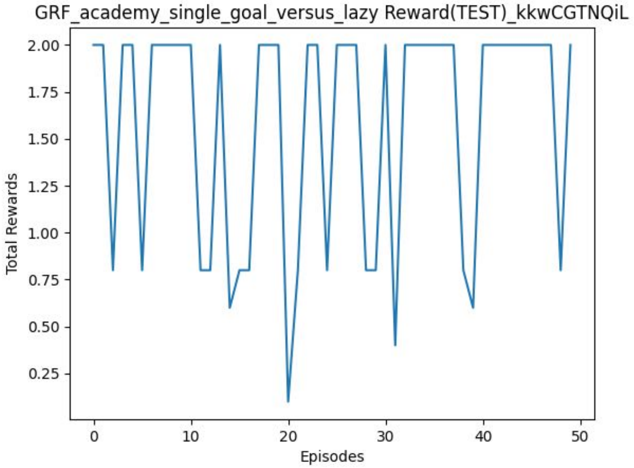




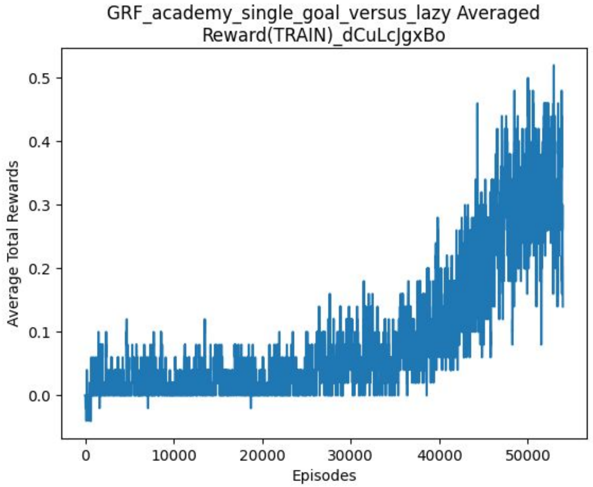


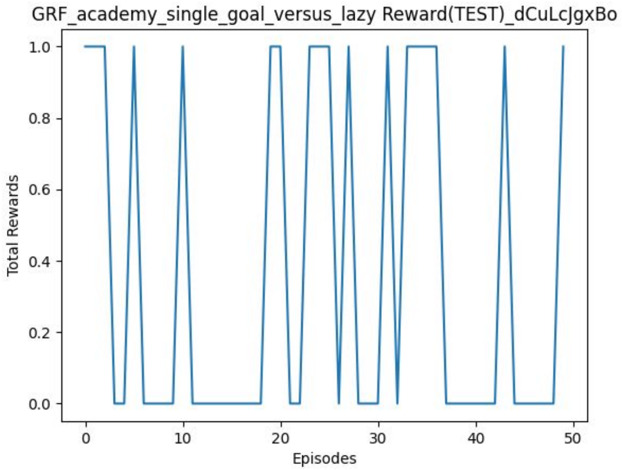
PPO - Dense



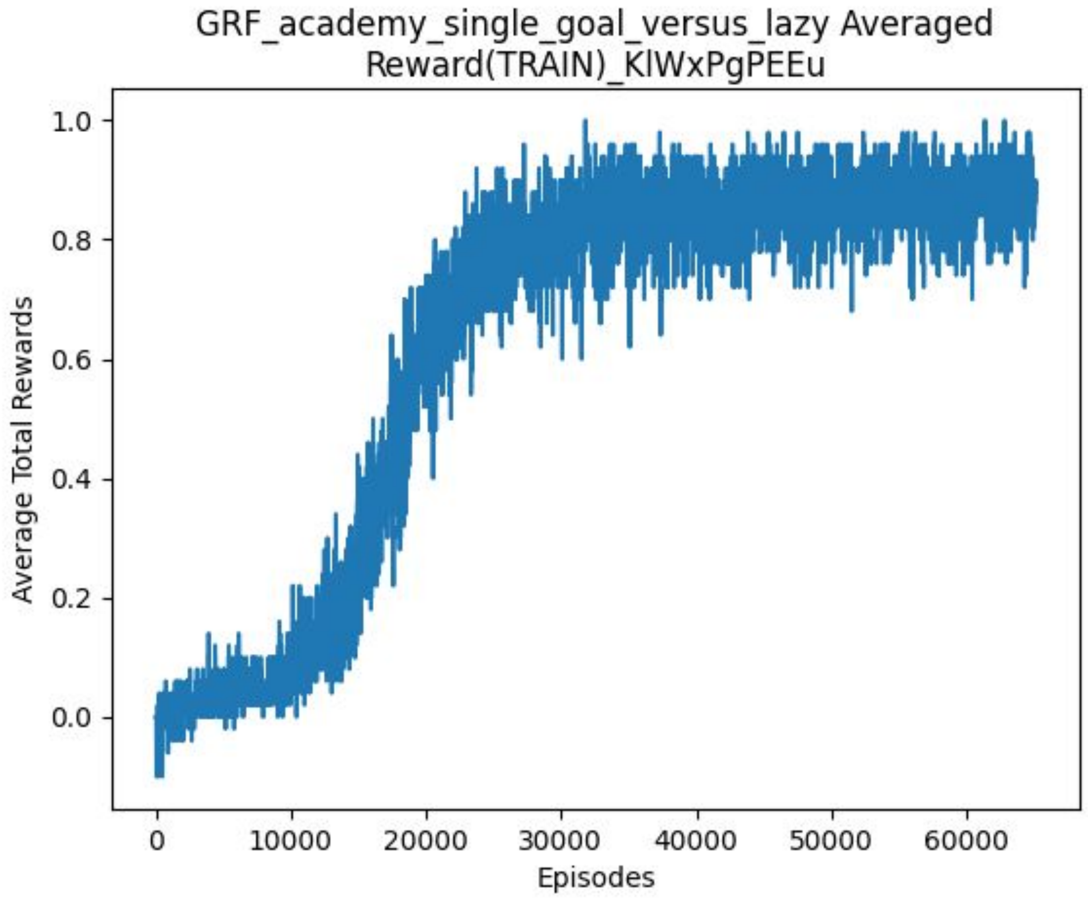


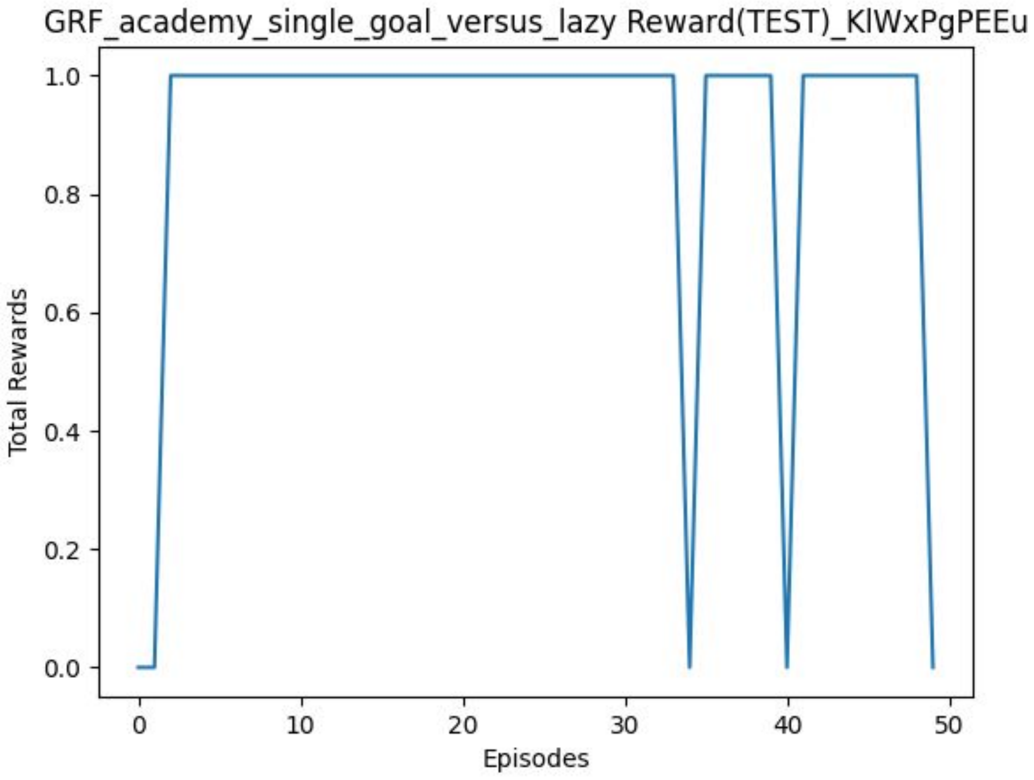
PPO - Sparse





PPO Sparse


GRF_academy_single_goal_versus_lazy Averaged Reward(TRAIN)_KlWxPgPEEu
Average Total Rewards
Episodes
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
50000
60000


GRF_academy_single_goal_versus_lazy Reward(TEST)_KlWxPgPEEu
Total Rewards
Episodes
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50

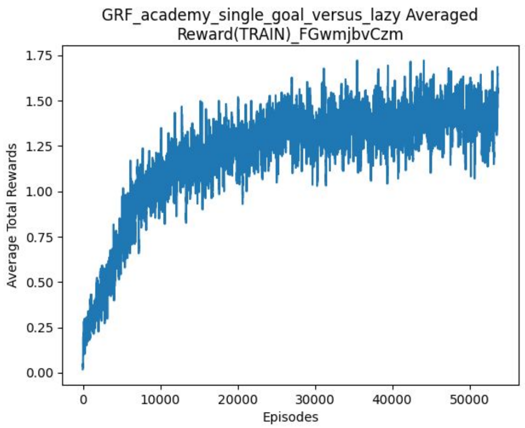
PPO - Sparse Visualization

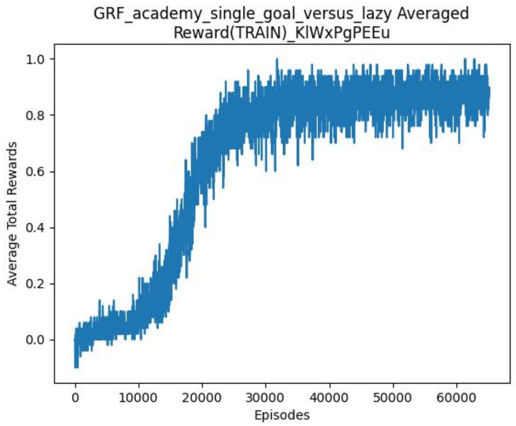
PPO - Dense Visualization

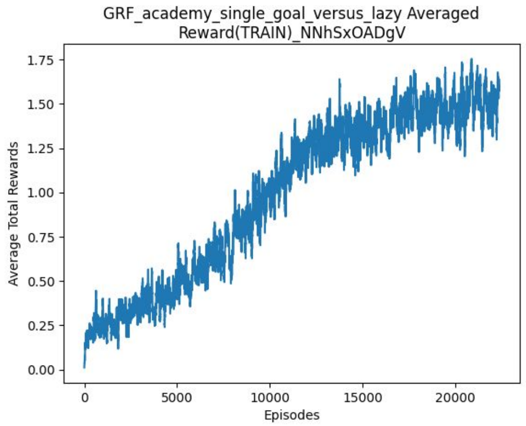
Sparse vs Dense Comparison

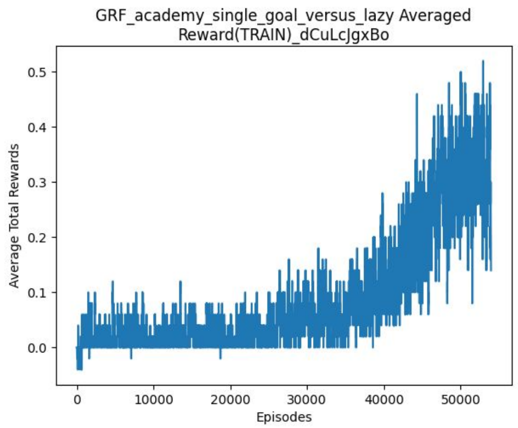
- The actor-critic agent performed well using the sparse reward structure.
- The dense reward structure occasionally confounded the agent into a local maximum in which it ran the ball to the opposing team’s goal line and out of bounds (corner kick).
- More training episodes were needed with the sparse reward structure, but final policy performance was better than that of the dense reward structure.

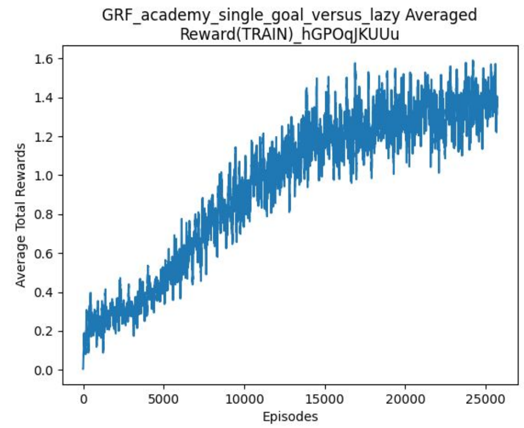
Hyperparameter Tuning

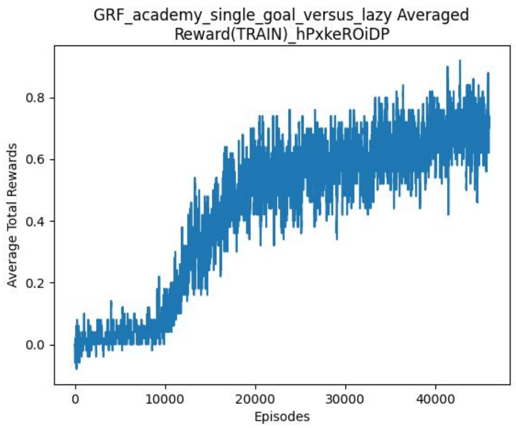


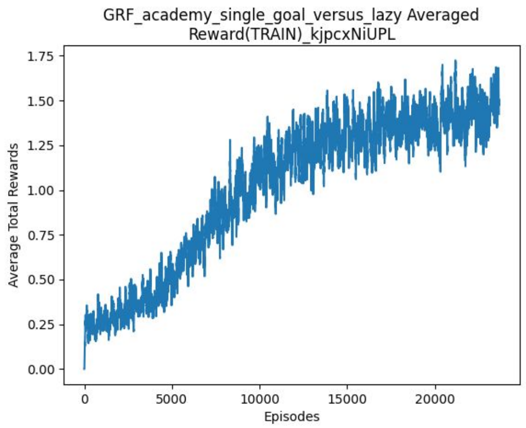


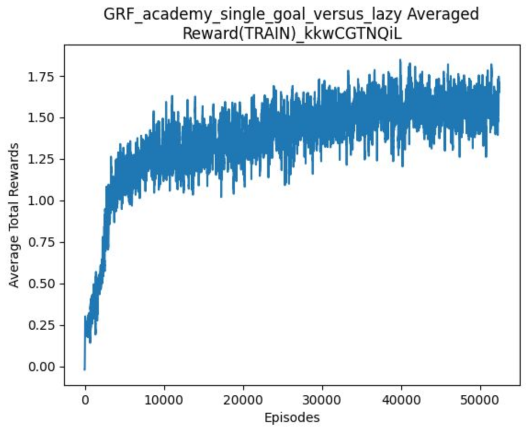












Future Work

- Implement an ensemble PPO algorithm with an aim to improve performance and stability of agent
- Implement our own reward structure to find the best fit for Google research football
- Solve Google Research Football against hard AI

References

- Google Research Football
- Environments
- Observations and Actions
- PPO
- Github Repository

Thank you