Multi - Agent Reinforcement Learning

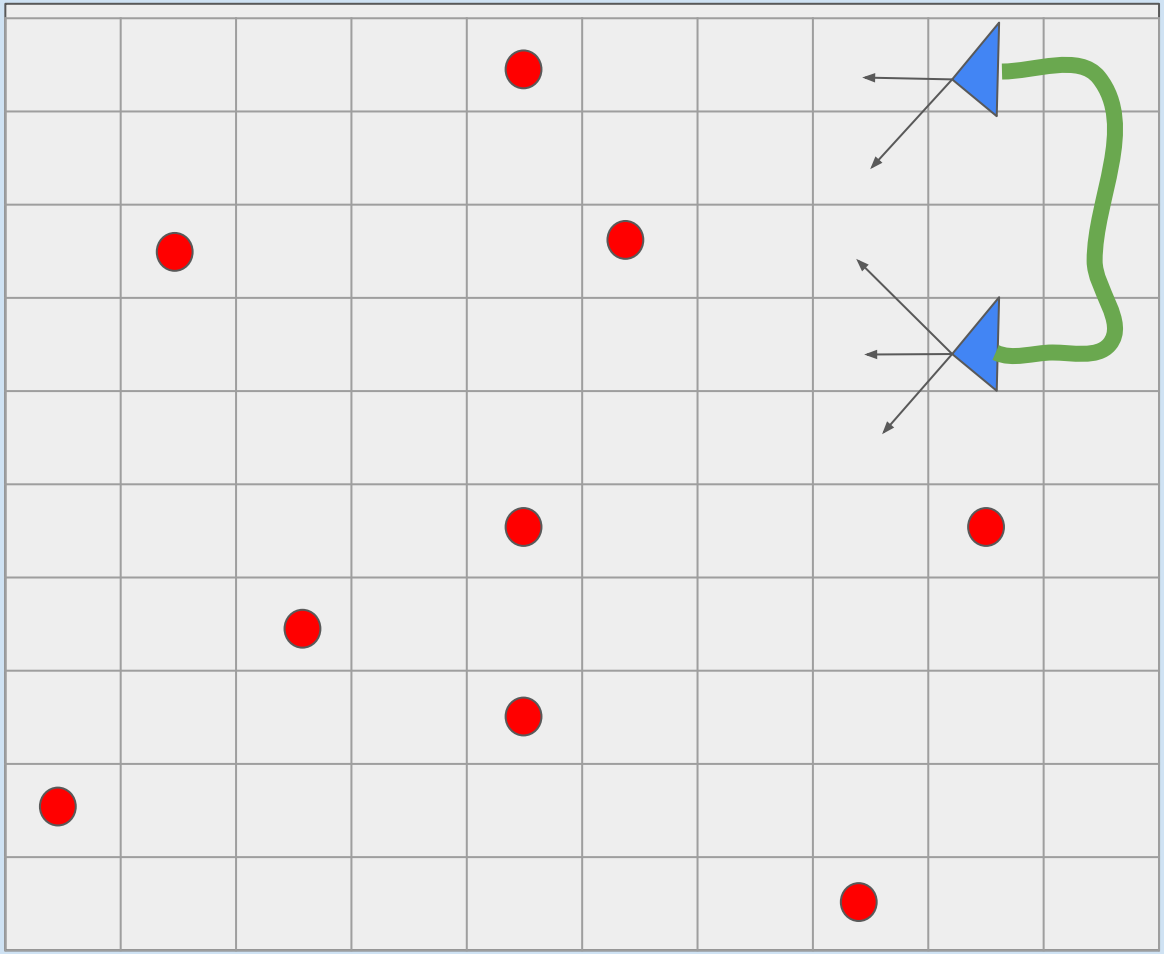
Semester Project

Multi-Agent Trash Collecting Framework for Lakes and Dams

Agamdeep Singh 20021
Nishant Chaudhari 21353
Karthik Nambiar 2320702

Problem description

- Objective: Create a Tethered Multi-Agent Trash Collecting Framework using MARL
- Environment: Represented as a 10x10 grid
  - 0 = empty
  - 1 = trash (red circle)
  - 2 = tether (green line)
  - 3 = 1st boat (blue triangle)
  - 4 = 2nd boat (blue triangle)
- Possible Actions: 9 actions (8 nearest neighbors + stay)



Proposed method - CNN (vision based)

- Instead of directly training the multi agent system in the grid world, a birds-eye-view perspective of the environment is made the input for the boats.
- Using CNNs as the feature extractor these boats perceive the environment and take appropriate actions.
- Extracted information is additionally fed:
  - Boat position
  - Active boat (which boat taking action)
  - Remaining trash: integer

Discussion

- Action Space: 3 → 9 (multi directional grid 3x3)
- Rewards:
  - Problem with previous rewards:
    - Discrete reward space → Hard to do gradient descent
    - Sparse rewards → Appears to be random
    - Noisy rewards → Sudden Jumps from negative to positive to negative, hard to discern what caused what
- Architecture: DQN with shared weights
- Sequential actions by boats ← 1 or 0 Neuron signifies which boat is acting

Training Parameters

Gamma - 0.95

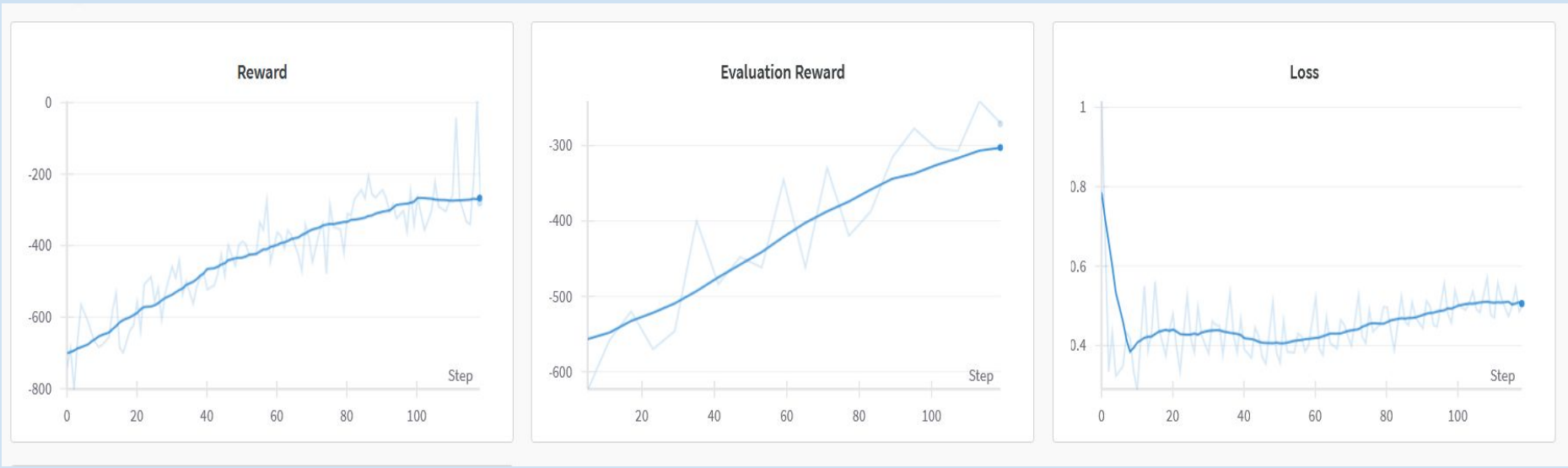
Eps Start - 1.0

Eps end - 0.05

Eps decay - 2000

No. of agents - 2

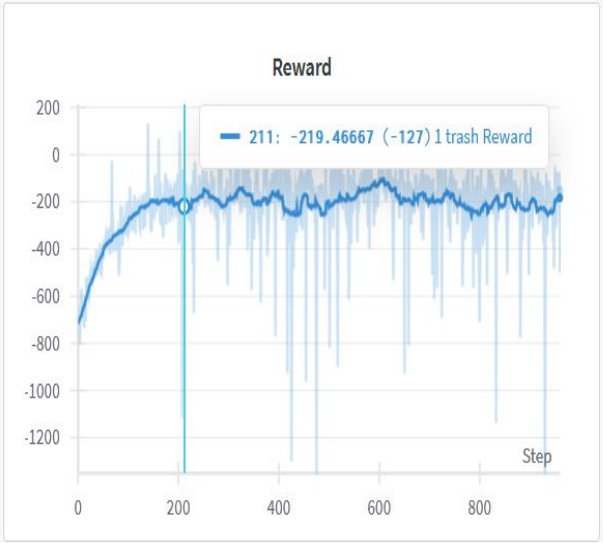
Initial Try

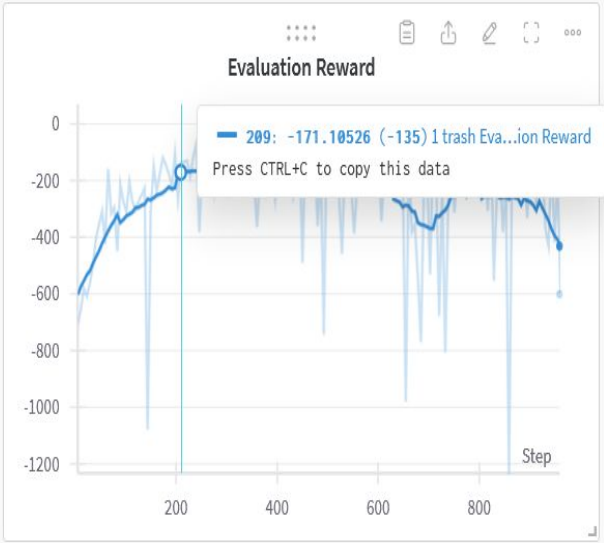

Reward
0
-200
-400
-600
-800
0
20
40
60
80
100
Step
Evaluation Reward
-300
-400
-500
-600
20
40
60
80
100
Step
Loss
1
0.8
0.6
0.4
0
20
40
60
80
100
Step

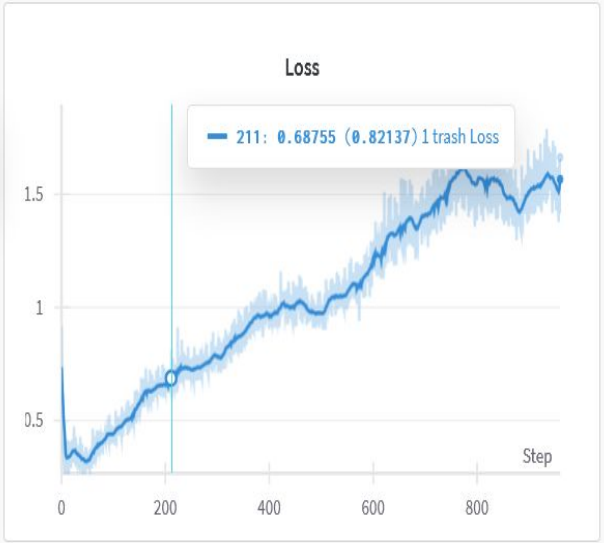
Training For 800 Episodes - Model Converged

Changed Parameter -

Grid = 13*13, Steps = 300, Trash = 5, TL = 5







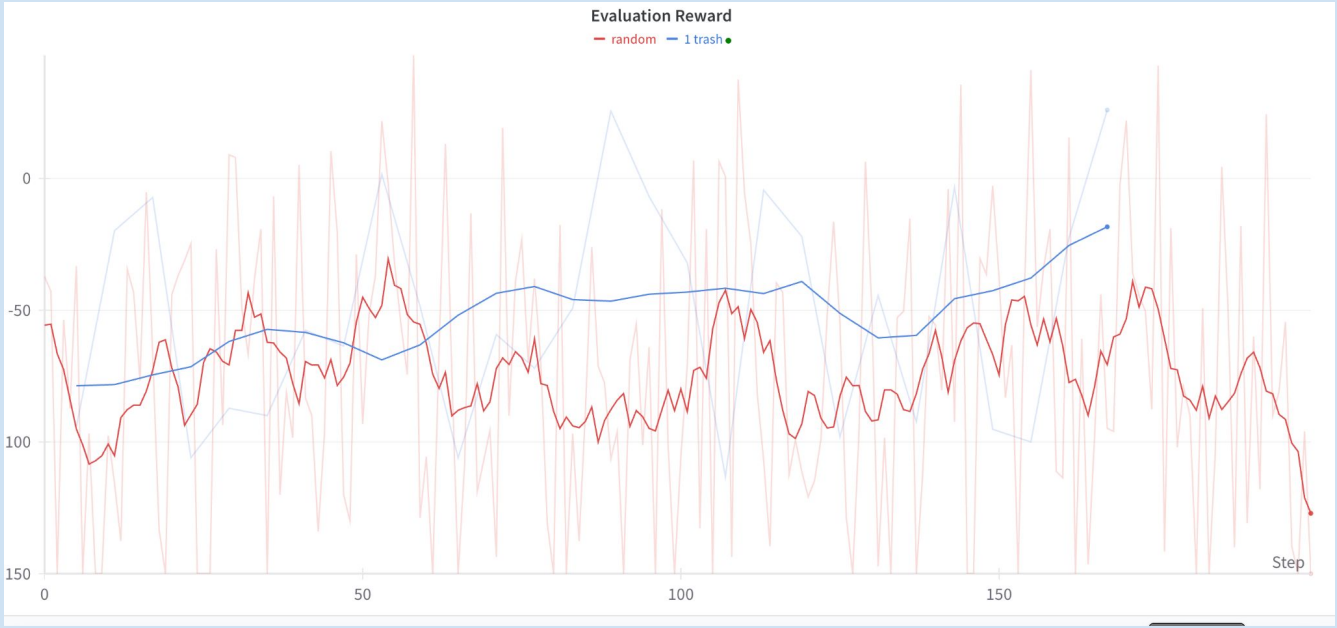
What didn’t work -

- Trash gain instance reward: Too sparse
- Invalid move penalty: model finds local optima in making legal moves only
- Proximity reward to closest trash: Model goes near trash and stands there

What worked -

- Make the reward surface smoother
- Rewards:
  - Time penalty: at every step
  - Trash gain reward: At every step
    - Models learns that getting trash faster results in higher compounding

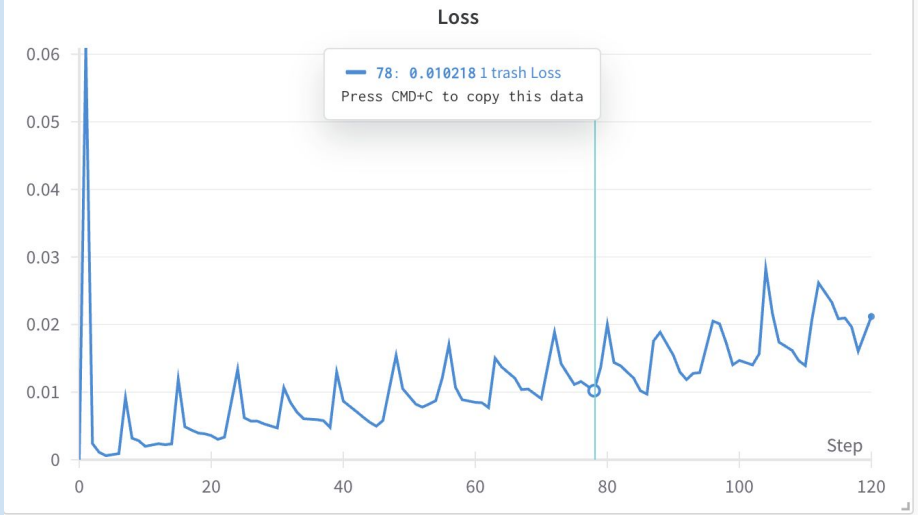
Evaluation

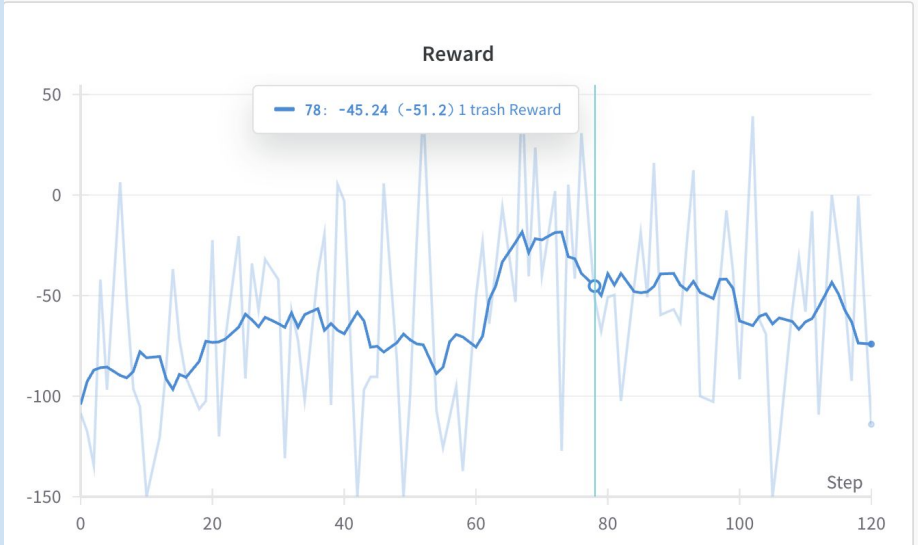


- 10x10 grid
- 5 trash
- Tether length: 4

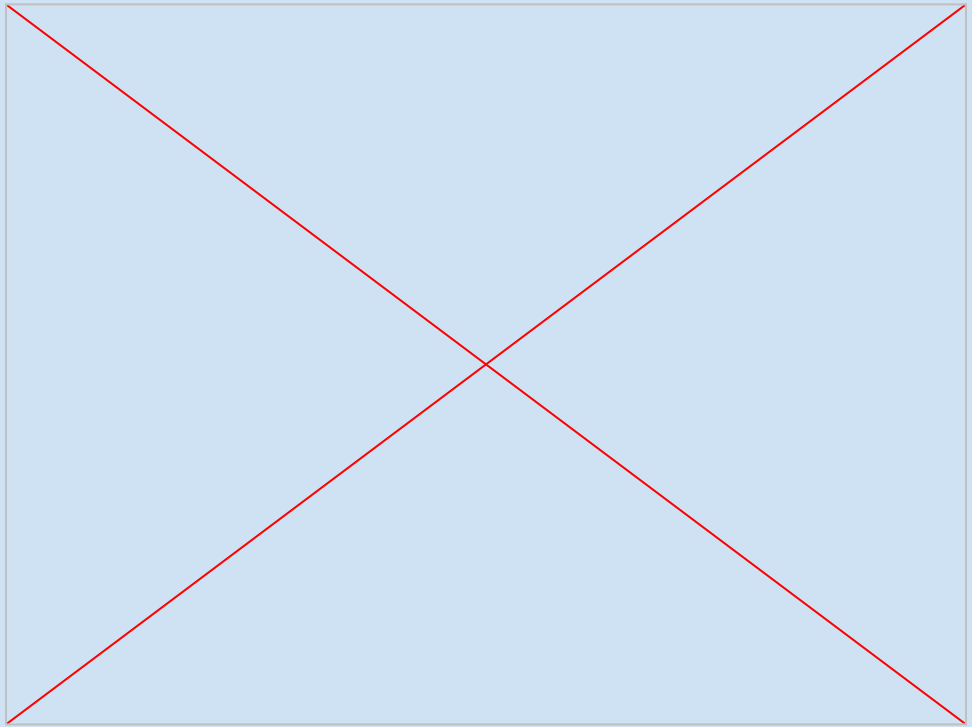
Still not converged

- Loss going up → Healthy training exploration is still going on
- Loss reflects change in weights per episode (target - policy)
- When loss stabilises → Policy has been trained





Random Actions



Learned Policy

Contribution -

Task	Person	Weightage
Environment Development	Karthik	Equal
Agent and training setup	Agamdeep	Equal
Parameter tuning and testing	Nishant	Equal

Thank you!