Reinforce With BaseLine Agent

By Sefunmi ASHIRU

Audio Based Environment

Legend:

X : Agent
O : Tree

Environment:

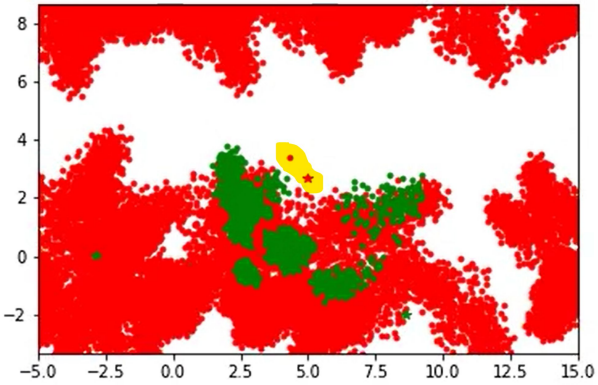
2D - audio input
Continuous
Partially Observable

Agent:

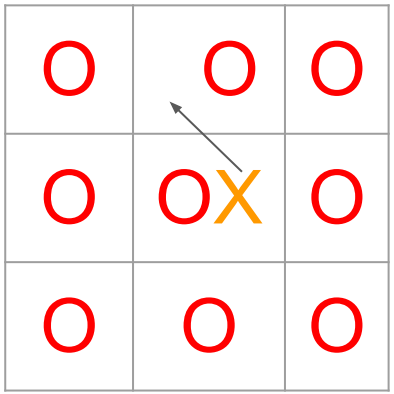
Reinforce with Baseline

GitHub:

https://github.com/michaeljgolds/sonar-rl-proj

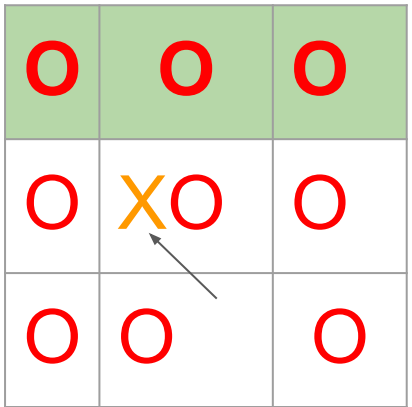


State t



Steps-Actions:
1-F, 2-R, 3-F

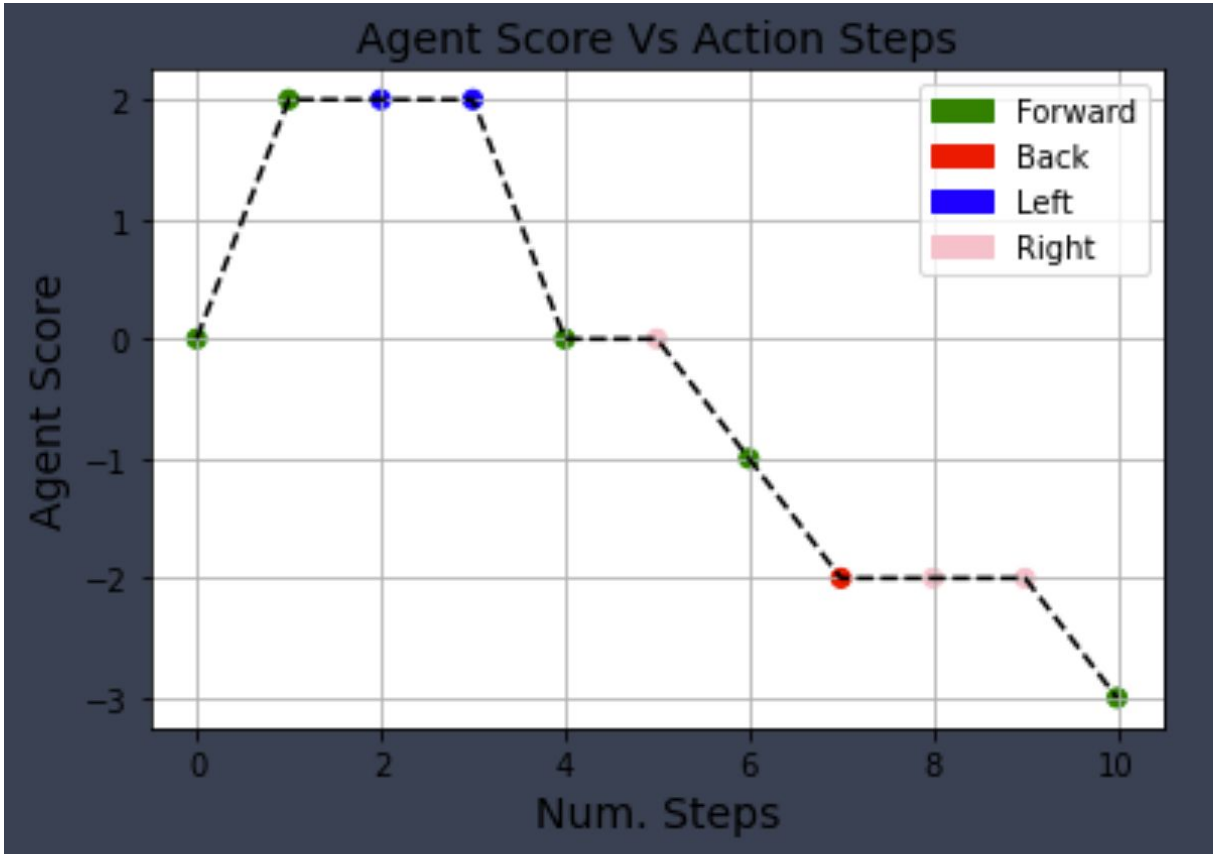
State t+3



Reinforce with Baseline

1. Initialize the policy parameter θ at random.
2. Generate one trajectory on policy π_θ: $S_1, A_1, R_2, S_2, A_2, \ldots, S_T$.
3. For t=1, 2, … , T:
    1. Estimate the the return G_t;
    2. Update policy parameters: $\theta \leftarrow \theta + \alpha\gamma^t G_t \nabla_\theta \ln \pi_\theta(A_t|S_t)$

Episode Tracking

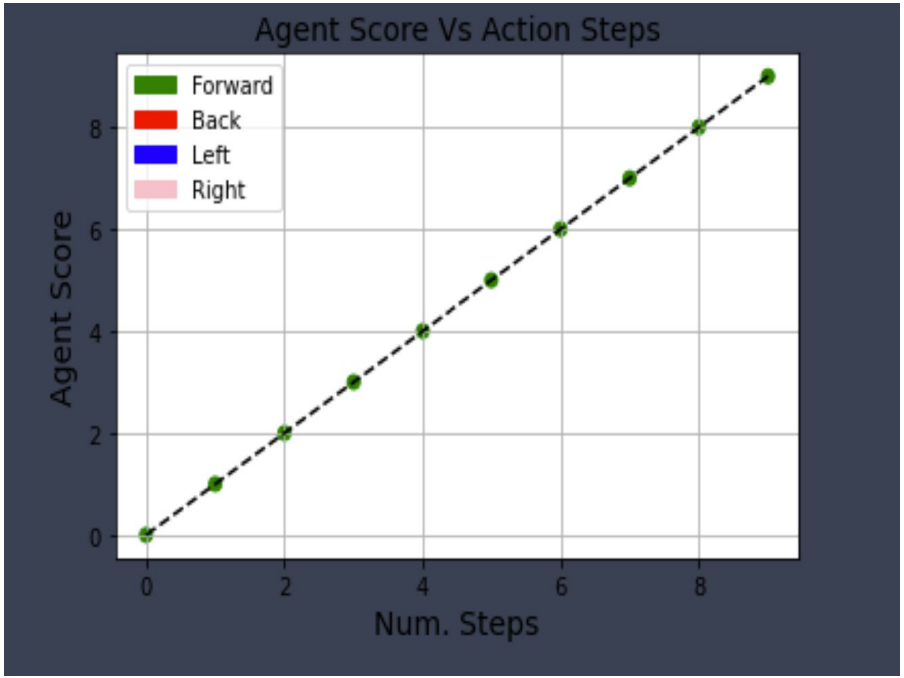


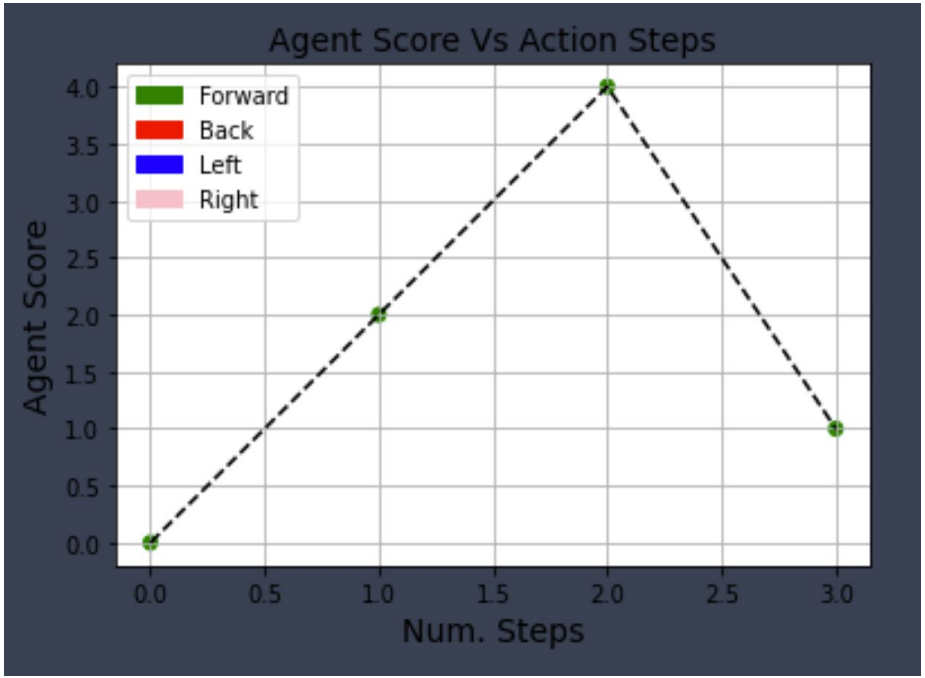
Training Progress

```
action:  R       score:    0.0  steps:    0.0
action:  D       score:   -1.0  steps:    1.0
action:  R       score:   -1.0  steps:    2.0
action:  L       score:   -1.0  steps:    3.0
action:  R       score:   -1.0  steps:    4.0
action:  L       score:   -1.0  steps:    5.0
action:  D       score:   -2.0  steps:    6.0
action:  U       score:   -3.0  steps:    7.0
action:  L       score:   -3.0  steps:    8.0
action:  U       score:   -1.0  steps:    9.0
ep:     19  score :   -1.0  steps : 9.0    average_score :   -3.5
action:  D       score:   -2.0  steps:    0.0
action:  R       score:   -2.0  steps:    1.0
action:  U       score:   -3.0  steps:    2.0
action:  R       score:   -3.0  steps:    3.0
action:  U       score:   -5.0  steps:    4.0
action:  L       score:   -5.0  steps:    5.0
action:  D       score:   -6.0  steps:    6.0
action:  D       score:   -7.0  steps:    7.0
action:  R       score:   -7.0  steps:    8.0
action:  D       score:   -5.0  steps:    9.0
ep:     20  score :   -5.0  steps : 9.0    average_score :   -3.5
action:  D       score:   -2.0  steps:    0.0
action:  L       score:   -2.0  steps:    1.0
action:  R       score:   -2.0  steps:    2.0
action:  U       score:    0.0  steps:    3.0
action:  R       score:    0.0  steps:    4.0
action:  U       score:   -1.0  steps:    5.0
action:  U       score:   -2.0  steps:    6.0
```

```
ep:   1764  score :    3.0  steps : 2.0    average_score :    3.0
action:  U       score:    1.0  steps:    0.0
action:  U       score:    2.0  steps:    1.0
ep:   1765  score :    2.0  steps : 1.0    average_score :    2.9
action:  U       score:    1.0  steps:    0.0
ep:   1766  score :    1.0  steps : 0.0    average_score :    2.9
action:  U       score:    1.0  steps:    0.0
ep:   1767  score :    1.0  steps : 0.0    average_score :    2.9
action:  U       score:    1.0  steps:    0.0
action:  U       score:    2.0  steps:    1.0
ep:   1768  score :    2.0  steps : 1.0    average_score :    2.8
action:  U       score:    1.0  steps:    0.0
action:  U       score:    2.0  steps:    1.0
action:  U       score:    3.0  steps:    2.0
action:  U       score:    4.0  steps:    3.0
action:  U       score:    5.0  steps:    4.0
action:  U       score:    6.0  steps:    5.0
action:  U       score:    7.0  steps:    6.0
action:  U       score:    8.0  steps:    7.0
action:  U       score:    9.0  steps:    8.0
action:  U       score:   10.0  steps:    9.0
ep:   1769  score :   10.0  steps : 9.0    average_score :    2.9
action:  U       score:    1.0  steps:    0.0
action:  U       score:    2.0  steps:    1.0
action:  U       score:    3.0  steps:    2.0
ep:   1770  score :    3.0  steps : 2.0    average_score :    2.9
action:  U       score:    1.0  steps:    0.0
```

Policy Learned





Improvements

- Issue: Agent seems to converge at shallow optima (Just move forward)
  - Neural Network : Currently just Relu 4 hidden Relu layers -> output classification 4
    - Add sigmoid for nonlinear
    - Decrease learning rate
    - Spectrogram (Qu: Do bats hear on the same logarithmic scale as humans?)
    - Output a sequence of actions (a0,a1,a2,a3,an) -> (r0 + r1 + r2 + r3 + rn)
    - CNN for feature extraction
  - Reward Shaping : Currently (N = 2, S = -2, E = -1, W = -1)
    - Attempt: (N = 1, S = -1, E = 0, W = 0) - longer training (GPU's ?)
  - Max Steps/Episode : Trained with max step 10 per episode
    - Incrementally increase max_steps in between saved weights training
  - Environment :
    - Reduce tree row gap & Increase gaps between tree trunks

Research Ideas

- Research audio data as input for RL.
- Reshape rewards.
  - Gain reward beacon, north edge o3
- Reshape policy gradient model towards deep RL.
  - Orientation information [input_audio, compass_heading] o1
  - Policy for multiple steps/path output (action1, action2, action3, action4) o2
  - Look into RNN - send past env_state/weights o4