31TH JULY 2024

MIGe Departiment
University of Studies of Trieste

Deep Reinforcement Learning methods for Google Research Football RL Environment

Student:
Gabriel Masella [SM3800048]

Exam project for:
Reinforcement Learning

Table of Contents:

1.

RL Environment

Brief introduction of the environment used in this project, what characterizes it and how it is described

1.a. Google Research Football (GRF) Environment

Google Research Football (GRF) is a reinforcement learning environment, created by the Brain Team of Google, where agents can be trained to play football in an advanced, physics-based 3D simulator.[1]

In the Reinforcement Learning context, that's the case where there is an **high observability** but **limited knowledge of the environment**.
So, we will look for **model-free algorithms.**

1.b State, Action and Reward

In this context, the term "**state**" refers to the **comprehensive data set returned by the environment in response to the completion of specific actions**. The state can be represented in one of three ways:

- **Pixels**:
  - It consists of a **1280 × 720 RGB image** corresponding to the rendered screen.
- **Super Mini Map (SMM):**
  - The SMM representation consists of **four 72 × 96 matrices**.
  - These matrices encode binary information about the home team, the away team, the ball, and the active player respectively.
- **Floats:**
  - The floats representation provides a compact encoding and consists of a **115-dimensional vector**, summarizing many aspects of the game, such as players coordinates, ball possession and direction, active player and game mode.[1]
  - **This is the one chosen for this experiment, because of its efficiency**

1.b State, Action and Reward

That’s the set of possible actions that a **single-agent** can choose from in this RL Environment:

- Standard **move actions (in 8 directions)**
- Different ways to **kick the ball** (**short** and **long passes**, **shooting**, **high passes**).
- Players can **sprint**, try to **intercept** the ball with a **slide tackle** or **dribble** if they possess the ball.
- **Moving** and **sprinting** actions are **sticky** and continue until an **explicit stop action** is performed.[1]

Top	Bottom	Left	Right
Top-Left	Top-Right	Bottom-Left	Bottom-Right
Short Pass	High Pass	Long Pass	Shot
Do-Nothing	Sliding	Dribble	Stop-Dribble
Sprint	Stop-Moving	Stop-Sprint	—

1.b State, Action and Reward

For every action made by the agent at each timestep, there is a corresponding reward

- Two reward functions, implemented by GRF: **SCORING** and **CHECKPOINT**.
  - **SCORING**: Each team obtains a **+1 reward when scoring a goal**, and a **-1 reward when conceding one to the opposing team**.
  - **CHECKPOINT:** The agent gets an **extra reward if they move the ball close to the opponent's goa**l. The agent's team gets an **extra reward of +0.1 to +1** when **they have the ball in each checkpoint region**. Checkpoint rewards are given once per episode.[1]

1.c Training Scenarios

The GRF environment has a **Football Academy** for researchers with limited processing resources.
It has 11 tasks of different difficulties, and a simple API lets users customize it.[1]
In this project **6 levels of training** are created to train and test our agent:

- *Level 0*: the Agent has to **search for the ball** (behind of him) and has to **score in an empty goal**.
- *Level 1*: the Agent, outside of the box, has to **score a goal**, but there is the **goalkeeper**.
- *Level 2*: the Agent has to score a goal but there is a **defender** near to him.
- *Level 3*: **2v1** where the agent is tagged by a defender and there is a **free team-mate**.
- *Level 4*: **3v2** situation with the agent tagged and another team-mate on the right tagged, the one on the left is free.
- *Level 5*: Complete simulation of **7v7 match** with **2-1-2-1 scheme.**

2. Solution Proposed

Introduction to Deep Reinforcement Learning and its advantages and a description of the Deep Q-Network algortim used in this project

2.a TD-Learning

Because of high observability and low knowledge of the environment, we resort to **TD-learning algorithm.**

TD learning learns **value function** $V(s)$ directly from experience, in a model-free, online, and fully incremental way.

TD learning is a prediction problem. The **update rule** is

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

The terms in parenthesis is known as ***Temporal Difference Error (TDE)*** and there are other kind of TDE like **SARSA**, **Exp. SARSA** or **Q-Learning.**[4]

2.a TD-Learning

Q-Learning, in particular, is an **off-policy control method** to find the **optimal policy**.

It learns **action-value function** $Q(s,a)$ and the **Bellman Equation** gives the following update rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

Although theoretically speaking, an agent based on Q learning **can memorize all state-action pairs from a table**, also known as a **Q table**.

But practically, it is computationally expensive when state and action space is enormous.

To deal with this problem, one can use a **function approximator** in order to approximate Q-values.[3]

2.b Deep Reinforcement Learning

With non-linear function we obtain **Deep Reinforcement Learning** methods when we use **deep neural networks** to **approximate** any of the following components of reinforcement learning:

- **Value Function** $\hat{v}(s;\theta)$
- **Action-Value Function** $\hat{q}(s,a;\theta)$
- **Policy Function** $\pi(s|a;\theta)$
- **Model** (Reward Function or State Transition Function)

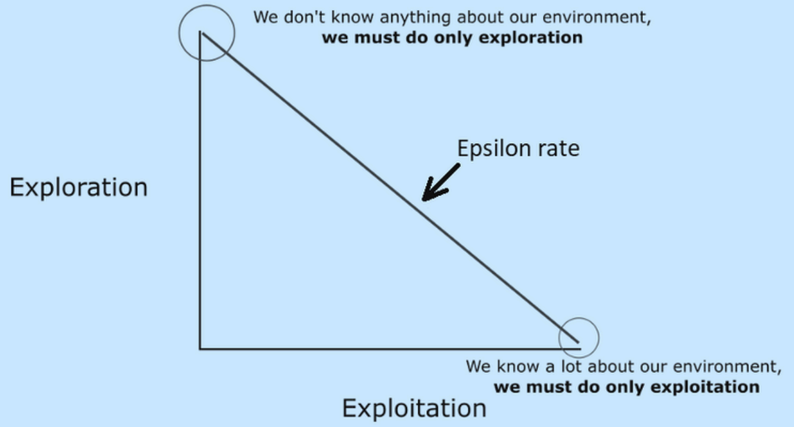
where θ are the network's weights.[4]

2.c Deep Q-Learnin Algorithm

Deep Q-Learning comes under reinforcement learning and is an **improvement** and **extension** of **Q-Learning** by introducing a ***Deep Neural Network for action-value function estimation***, ***Replay Buffer*** and an ***Exploration Method***.

With ***experience replay*** the **agent's experiences are stored** at each time-step $e_t = (s_t, a_t, r_t, s_{t+1})$ into a *replay memory* and random samples of experience tuples are drawn out during the Q-learning updates.

After the execution of experience replay, the agent selects and executes an action according to an **ε-greedy policy**, designed to achieve a **balance between exploration and exploitation**.[2]



Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

2.c Deep Q-Learnin Algorithm

This approach has **several advantages** over **standard online Q-learning**:

- each **step of experience is potentially used in many weight updates**, which allows for greater data efficiency.
- learning directly from consecutive samples is inefficient, due to the strong correlations between the samples, **randomizing the samples breaks these correlations and therefore reduces the variance of the updates.**
- By using experience replay **the behavior distribution is averaged over many of its previous states**, smoothing out learning and avoiding oscillations or divergence in the parameters. [2]

3.

Implementation and Training

Explanation of the designed architecture, hyper-parameters chosen and training procedure

3. Implementation and Training

The **Deep Q-Network**, developed for this project, comprises **three layers of 128 neurons** each, **a maximum size of 10,000 tuplets as replay memory**, and a **linear decay of the epsilon of 0.005**.

As an **off-policy method**, during the inner loop of the algorithm, **Q-learning updates are applied with minibatch updates**, with a **batch size of 256**.

```
##########################################
# Hyperparameters
##########################################

lr = 0.00001                # learning rate
gamma = 0.99                # discount factor
batch_size = 256            # batch size

epsilon = 1.0               # starting value of epsilon
epsilon_decay = 0.005       # decay rate of epsilon
epsilon_min = 0.1           # minimum value of epsilon

max_size = 10000            # max size of the replay buffer

fc1_dims=128                # number of neurons in the first layer
fc2_dims=128                # number of neurons in the second layer
fc3_dims=128                # number of neurons in the third layer
```

3. Implementation and Training

The **agent**, designated in the code as CR7*, **is trained through six training scenarios** of increasing difficulty.

The initial five levels have a maximum of **400 time steps**; however, this can be terminated prematurely if the objective is attained or the ball is removed from play.
For the 7v7 match simulation, there are **3,000 frames.**

Each scenario is comprised of **1,000 episodes** of training, followed by **100 episodes** of testing to assess the agent's learning outcomes.
The exception to this is the final level, where only 100 matches are used for training and 10 for testing, due to the computational demands of the task.

*= nickname of the famous player Cristiano Ronaldo, known for his great achievements through hard training

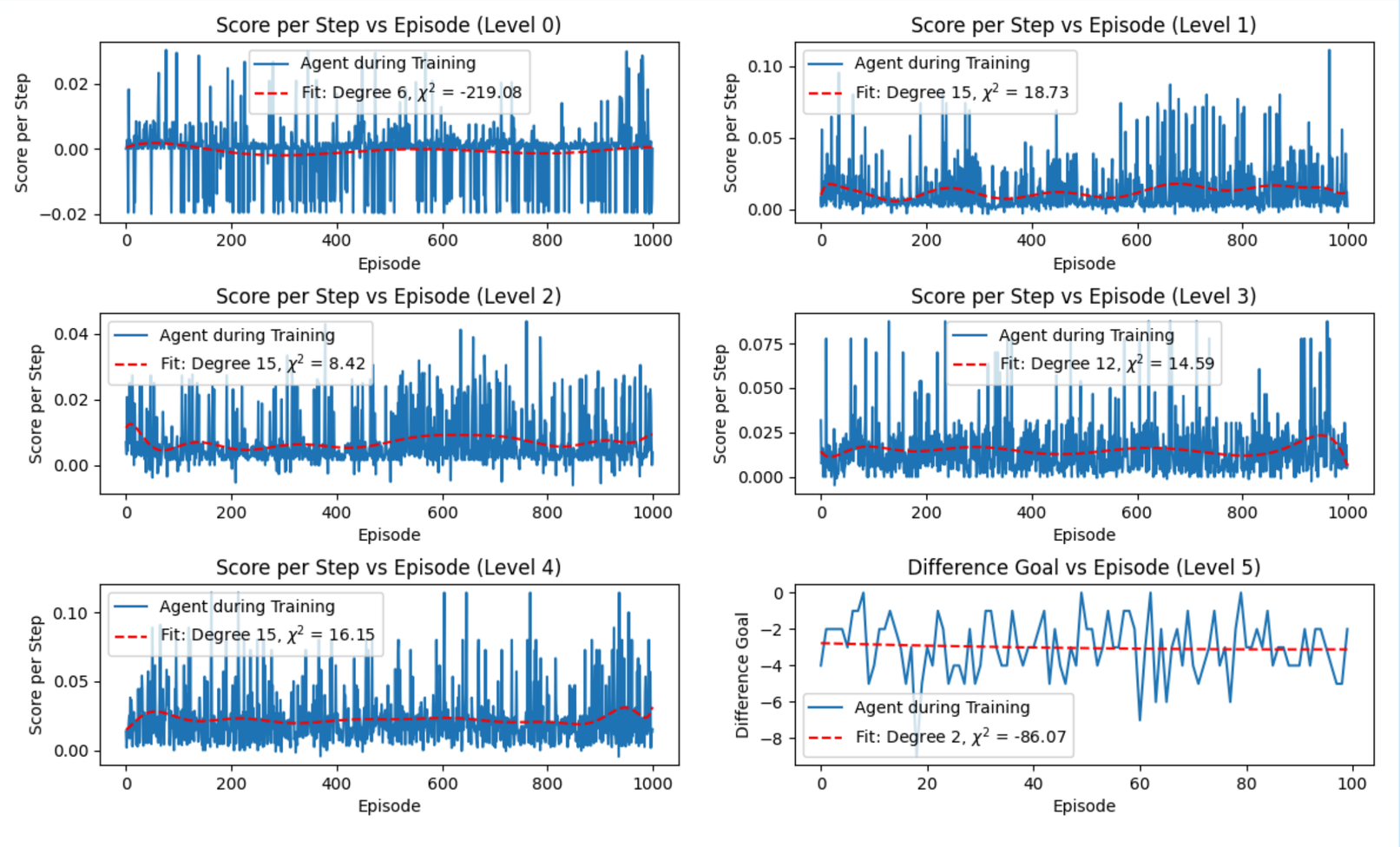
4.

Test Results

How the CR7 agent performed? Does it learn an optimal policy to solve different scenarios?

4. Test Results

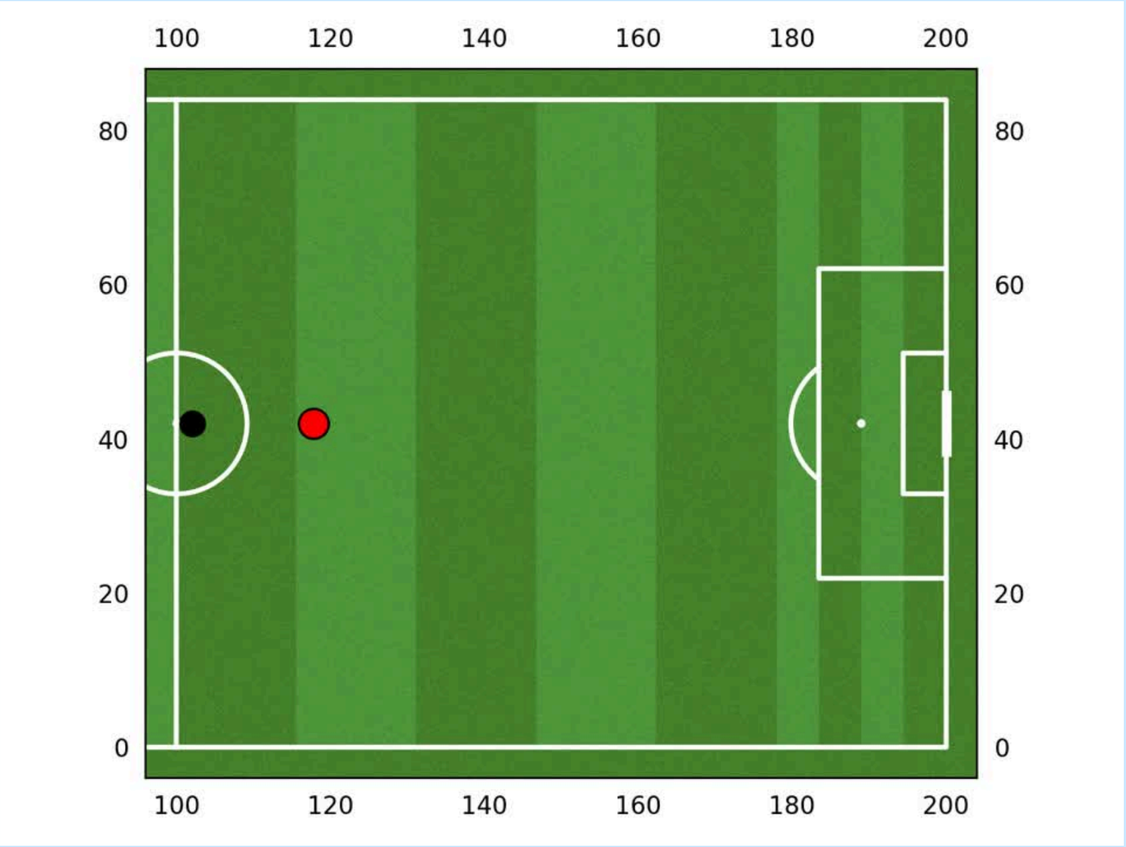
Prior to examining the knowledge acquired by agent CR7, it is instructive to consider the following graph, which illustrates the evolution of the score ratio on # step as a function of training level.



4. Test Results

The following visualization illustrates the policy that he acquired during the training. It is not the optimal policy, but he was able to identify a method for achieving his tasks.

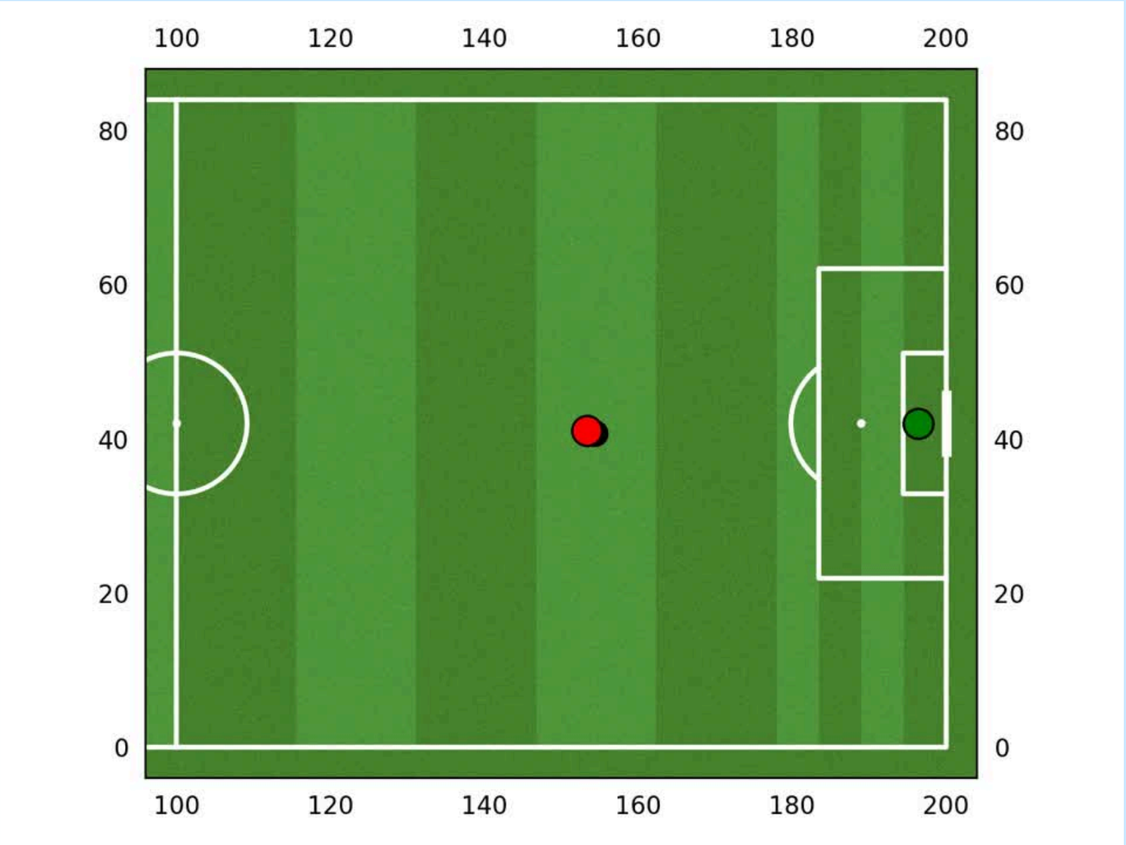
At the level 0, he was taught to locate the ball independently and to shoot from any position he chooses.
He is therefore able to score successfully.



4. Test Results

The following visualization illustrates the policy that he acquired during the training. It is not the optimal policy, but he was able to identify a method for achieving his tasks.

In contrast, at Level 1, he did not learn the optimal method for scoring.
Instead, he developed his own approach: attracting the goalkeeper to such an degree that the goal is unguarded, allowing him to score with relative ease.

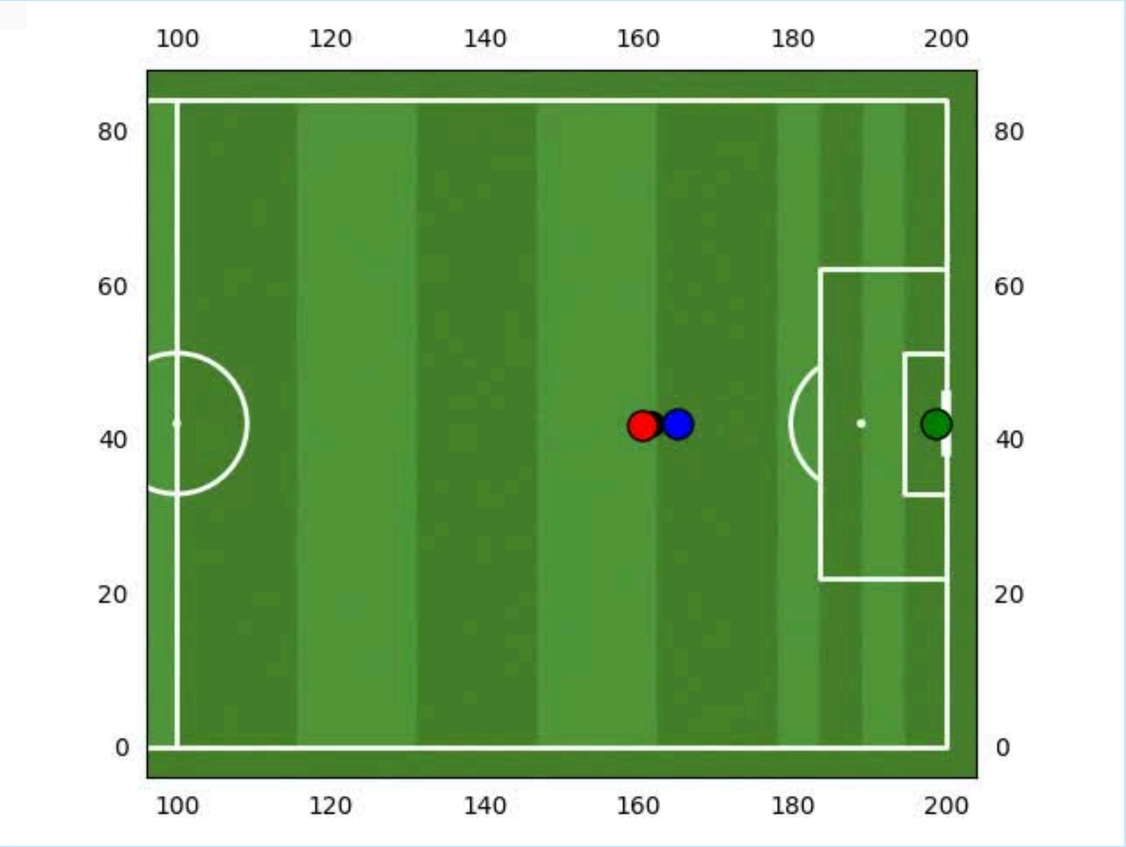


4. Test Results

The following visualization illustrates the policy that he acquired during the training. It is not the optimal policy, but he was able to identify a method for achieving his tasks.

From level 2 onward, the agent faces increasing challenges in adapting the previously identified policy to achieve the desired outcome.

In another training session, the agent employs a novel tactic, recognizing the potential of shooting from a distance to capitalize on the goalkeeper's rebound and subsequently score.



5.

Further Improvements

How the agent could do better, where the flaws are, and what to improve in the model.

5. Further Improvements

It is evident that the model is not operating at its optimal level for a number of reasons. However, there are numerous potential approaches that could be taken to enhance the performance of the agent:

- The implementation of a target network in the algorithm could serve to stabilize the learning process.
- A hyperparametrization-tuning could be performed.
- The use of different state representations, such as pixels, and the subsequent use of a convolutional neural network (as employed by Google in the original paper) could also be considered.
- Uniform sampling gives equal importance to all transitions in the replay memory. However, a more sophisticated sampling strategy might emphasize transitions from which we can learn the most, similar to Prioritized Experience Replay.

References

1. Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. (2020). **Google Research Football: A Novel Reinforcement Learning Environment.**
2. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. (2013). **Playing Atari with Deep Reinforcement Learning.**
3. Gupta, K., Dahiya, N., Dabas, M., and Pushparaj, P. (2022). **Playing football game using AI agents**. In *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)* (pp. 84-88).
4. Yuxi Li. (2018). **Deep Reinforcement Learning: An Overview.**

2.a TD-Learning

Discount factor $\gamma \in [0, 1]$ is used to **penalize future rewards**. There are many reasons to employ the discount factor in the equation. The most important among all is as follows:

- High level of uncertainty of future rewards
- Future rewards are not as beneficial as immediate rewards
- Acts as a mathematical convenience
- There is no need to worry about the infinite loops in the state transition graph

4. Test Results

The following table presents the test accuracy of the experiment conducted on agent CR7. For the initial five levels, we provide the number of times the agent scored goals as a percentage. For the final level, we present the average goal difference in the test matches.

Training Scenario	Test Accuracy:
Level 0	24%
Level 1	6%
Level 2	3%
Level 3	0%
Level 4	0%
Level 5	-3.01±1.61

B