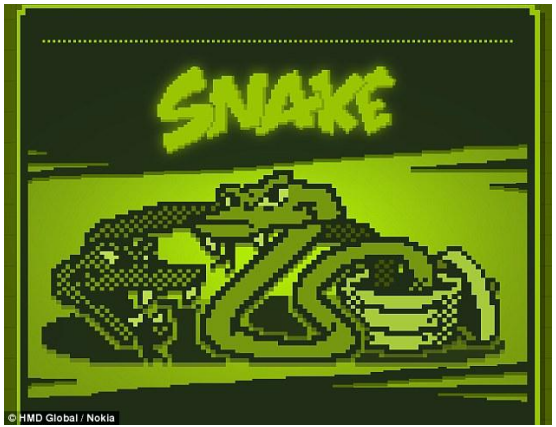
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DL@NOVA: 60055 59241 67476

Snake game



Team identification

Name 1: Ricardo Bernardino

Number 1: 67476

Name 2: José Diogo Romano

Number 2: 59241

Name 3: João Lopes

Number 3: 60055

Name 4:

Number 4:

Final private score:

Leaderboard private ranking:

Data analysis and preprocessing

The overall analysis and preprocessing of this assignment relied majorly on understanding how a DQN works specially in relation to the Snake Game. When should the snake learn, when should the target network be updated, how to balance exploration and exploitation, and how to preprocess the game frames to feed into the neural network efficiently etc.

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Markov Decision Process

Action space:

1. Move left
2. Move straight (no change in direction)
3. Move right

State space:

1. Positions of the snake's head and body segments.
2. Location of the apple.
3. Direction of movement of the snake's head.

Reward structure (if applicable):

1. Eating an apple: When the snake eats an apple, it receives a positive reward.
2. Crashing into itself or the game boundary: If the snake collides with itself or the game boundary, it receives a negative reward.
3. Surviving without eating an apple: The snake receives a small negative reward for each time step it survives without eating an apple.

DQN architecture

3 Linear Layer NN - Neural network with 3 fully connected layers. The input layer accepts observations of the environment, which are flattened to a one-dimensional tensor. The first two hidden layers have 128 neurons each and use the ReLU activation function. The output layer produces Q-values for each possible action in the environment.

Exploration strategies

Epsilon Greedy Strategy - Governs the agent's decision-making process during both training and gameplay. At the start, the agent explores the environment randomly with a probability of 1, represented by the epsilon parameter. At each step, it chooses between exploration and exploitation: with a probability epsilon, it selects a random action to discover new strategies, and with a probability of (1 - epsilon), it exploits its learned knowledge by choosing the action with the highest Q-value on the Q-Table. As the agent gains more experience with time, the focus is shifted from exploration to exploitation.

DQN hyperparameter tuning

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

1. Batch Size (BATCH_SIZE): Determines the number of transitions sampled from the replay buffer during each training iteration.
2. Discount Factor (gamma): Controls the importance of future rewards in the Q-value updates.
3. Exploration Rate (epsilon): Balances exploration and exploitation by determining the probability of selecting a random action versus the one with the highest Q-value.
4. Epsilon Decay (eps_decay): Controls the rate at which epsilon decays over time.
5. Replay Memory Capacity (replay_memory): Specifies the capacity of the replay memory buffer, which stores past experiences for training.
6. Target Network Update Rate (TAU): Defines the rate at which the weights of the target network are updated with the weights of the policy network.
7. Learning Rate (learning_rate): Determines the step size used by the AdamW optimizer during gradient descent.

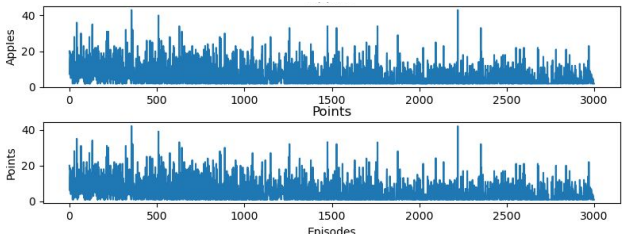
Ablation study: Comparison of DQN vs. DQN-ER

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

When we compare the performance between simply using DQN and incorporating Experience replay on the DQN, we can clearly see that:

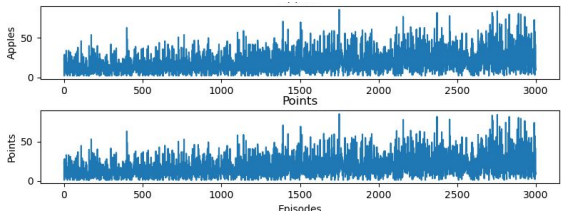
- The number of apples eaten by episode increases significantly, with a lot more apples per episode on later episodes of learning;
- By consequence, the score per episode and the average score both increase.

DQN



With this analysis, we can conclude that by incorporating ER the performance increase the extra computing power needed for the task

DQN-ER



Ablation study: Comparison of DQN vs. DQN-TN

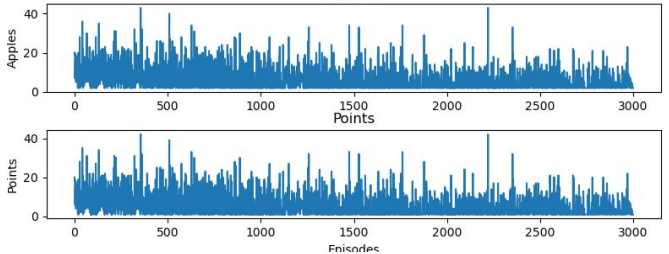
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

When comparing the performance of a standard DQN with a DQN that incorporates a Target Network (DQN-TN), it is evident that:

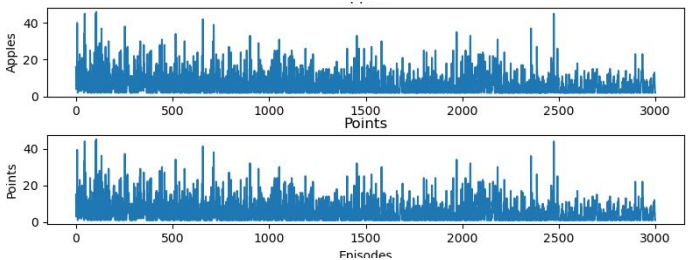
- While the performance improvement is not as dramatic as seen with Experience Replay (ER), the scores demonstrate increased stability when using a DQN-TN.
- Although there is a slight decrease in the scores, they tend to cluster more closely around the average, indicating a more consistent performance.

From this analysis, we can conclude that incorporating a Target Network into our DQN primarily enhances the stability of the scores, rather than significantly boosting the overall performance of the DQN.

DQN



DQN-TN



Ablation study: Comparison of DQN vs. DQN-ER-TN

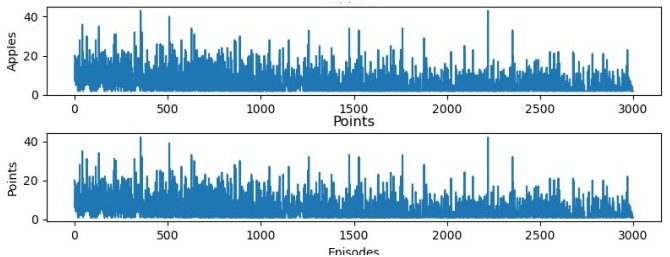
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

When comparing a simple DQN with a DQN augmented with Experience Replay (ER) and Target Network (TN), the results are striking:

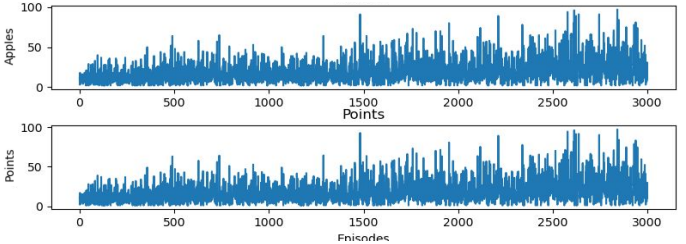
- The scores exhibit a remarkable improvement, nearly doubling on average. This surge in performance can be attributed to the synergistic effects of both ER and TN.
- Experience Replay enhances learning by breaking temporal correlations in the data, enabling the network to train on a more diverse set of experiences.
- This, combined with the stability provided by the Target Network, fosters a more consistent learning process.

The utilization of ER, especially when coupled with TN, significantly boosts the score, underscoring the effectiveness of these enhancements in augmenting the capabilities of the DQN framework. Overall, the incorporation of ER and TN not only elevates performance but also fosters more consistent and efficient learning dynamics, showcasing the potential of advanced techniques in RL.

DQN



DQN-ER-T



NOVA SCHOOL OF SCIENCE & TECHNOLOGY

What went wrong

Firstly, we are sad that we had a very small amount of time to improve our assignment, as well as investigate ways to improve it in itself, due to the overcumbering tests and assignments from other courses, which in this semester proved itself to be a very complex situation to conciliate both in the student side but also from the CP (Pedagogic Commission).

Given that, we were still able to develop a fairly simple model with a fairly simple DQN, that outputs (bad, but reasonable) results.

Regardless, we did our best, just like we did before on AA (Aprendizagem Automática, the previous course), with these limited time resources and exterior complications..

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

What went great

This project proved to be more captivating than many others in both AP and AA. We found the theoretical concepts engaging and enjoyed the hands-on application, particularly because the project involved something as relatable and enjoyable as the classic snake game. It offered a refreshing blend of theory and practical implementation, making the learning experience exceptionally enjoyable.