Laboratório 12 - Deep Q-Learning

Marcelo Buga Martins da Silva

CT-213 - Professor Marcos Ricardo Omena de Albuquerque Máximo

12/07/2021

1 Introdução

O propósito desse laboratório foi resolver o problema de Mountain Car usando o algoritmo seminal Deep Q-Learning de Deep Reinforcement Learning.

Para esse problema, o carro precisa subir ao topo da montanha mais a direita do espaço, podendo, a cada instante, empurrar para a esquerda, para a direita ou não fazer nada. O objetivo da rede neural é aprender uma política eficiente para a tomada de ações em posições e velocidades distintas garantindo que o carro consiga subir a montanha em uma quantidade significativa de vezes.

Para a implementação do laboratório, foi utilizado o *framework* Keras para a construção da rede neural. Mais detalhes acerca da implementação, assim como os resultados do treinamento da rede e da aplicação da política final após o treinamento são dados a seguir.

2 Implementação

A implementação dos algoritmos consistiu, primeiro, na implementação da rede neural utilizando o Keras. Foram criadas três camadas densas, as duas primeiras de 24 neurônios e uma camada de saída para cada ação

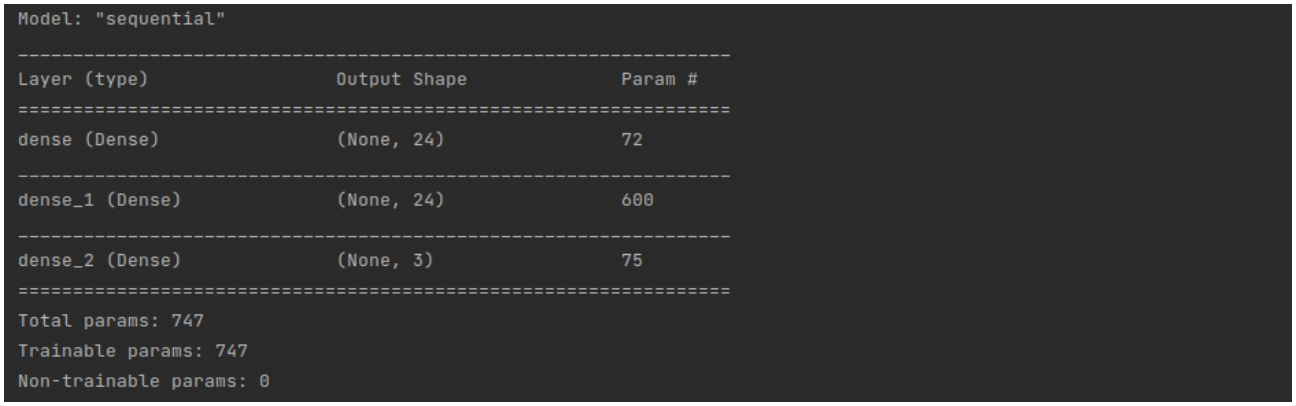

Figura 1: Sumário da rede neural implementada

possível (no caso desse laboratório, três neurônios). Utilizando a função *summary* do *model* do Keras, foi obtido o resultado apresentado na Figura 1:

Em seguida, implementou-se uma escolha de ação ε-*greedy*, isto é, toma-se a melhor ação dada pela rede neural com probabilidade $1-\varepsilon$ e uma ação aleatória com probabilidade ε, garantindo *exploitation* para encontrar a melhor política durante o treinamento. Por fim, configurou-se o sistema de recompensas, de forma que, em vez de recompensar somente pelo tempo gasto até chegar ao objetivo (ou extrapolar o tempo máximo), o algoritmo recompense mais políticas que levem o carro mais longe de sua posição inicial e tenham maior velocidade. Por fim, deu-se uma grande recompensa caso o carro atinja a posição alvo, bisando uma convergência mais rápida do treinamento.

3 Treinamento

O treinamento da rede neural foi realizado após 300 iterações iniciais. Embora o gráfico de convergência parecesse bons, como na Figura 2, quando foi devidamente avaliada a política, o resultado foi próximo de 50% de sucesso. Assim, mais algumas iterações de treinamento foram realizadas para atingir uma melhor política. A Figura 3 representa o resultado do treinamento posterior.

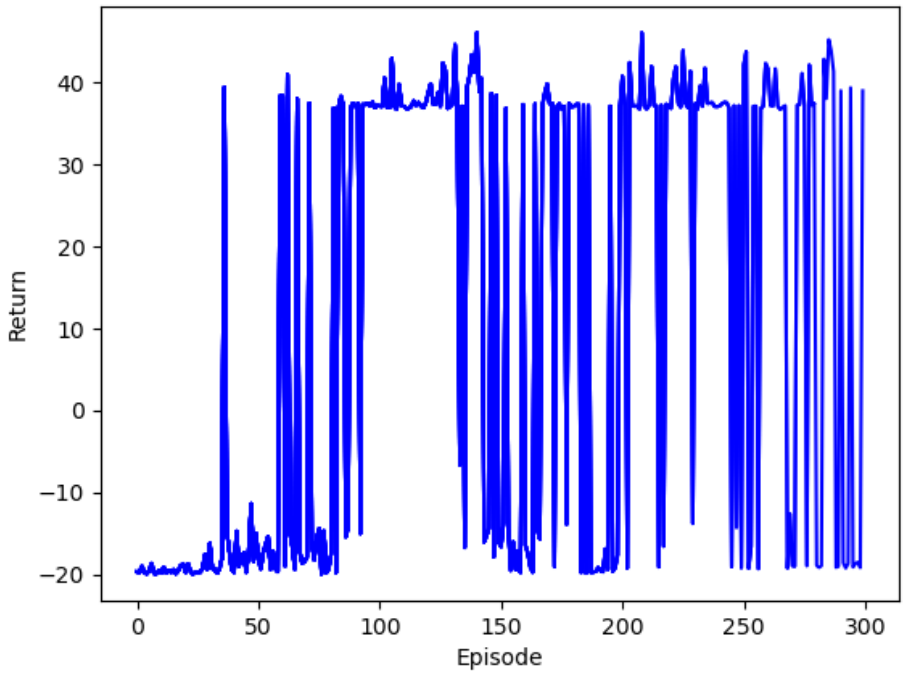

Figura 2: Treinamento inicial de 300 iterações

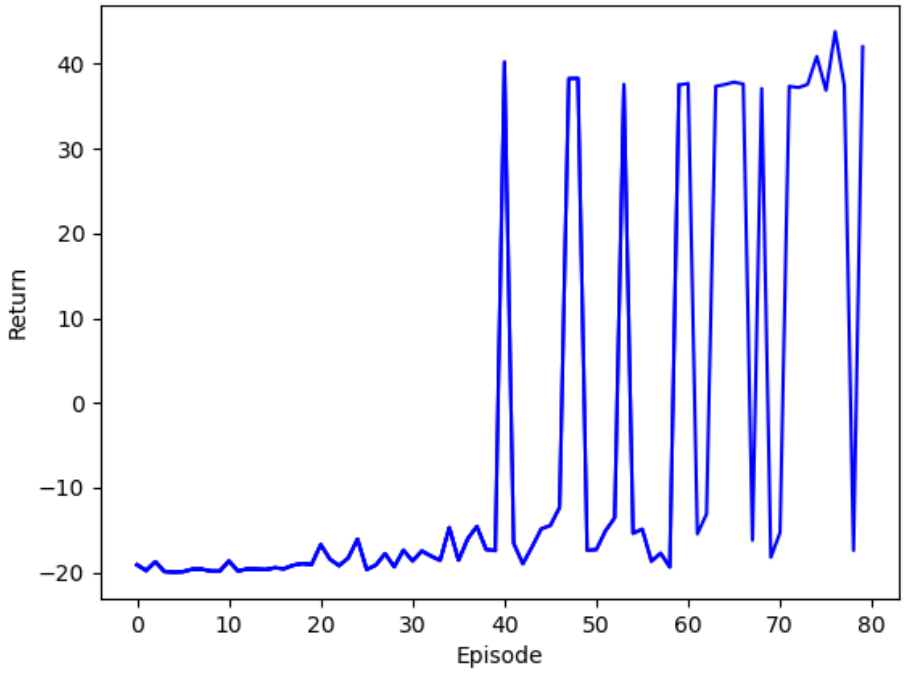

Figura 3: Refinamento do treinamento

Pode-se perceber que o treinamento é muito sensível a pequenas variações e, como evidente na Figura 2, o treinamento sucessivo pode piorar a política, que estava muito boa próxima da centésima iteração. Foi necessário interromper o treinamento quando as políticas passaram a ter bons resultados na segunda vez para garantir o sucesso da avaliação de política.

4 Avaliação de política

Após o treinamento, foi avaliada a política com 30 tentativas de subida do morro, tendo sido todas elas bem sucedidas. As Figuras 4 e 5 mostram a recompensa de cada uma das tentativas e a política adotada para cada par posição-velocidade.

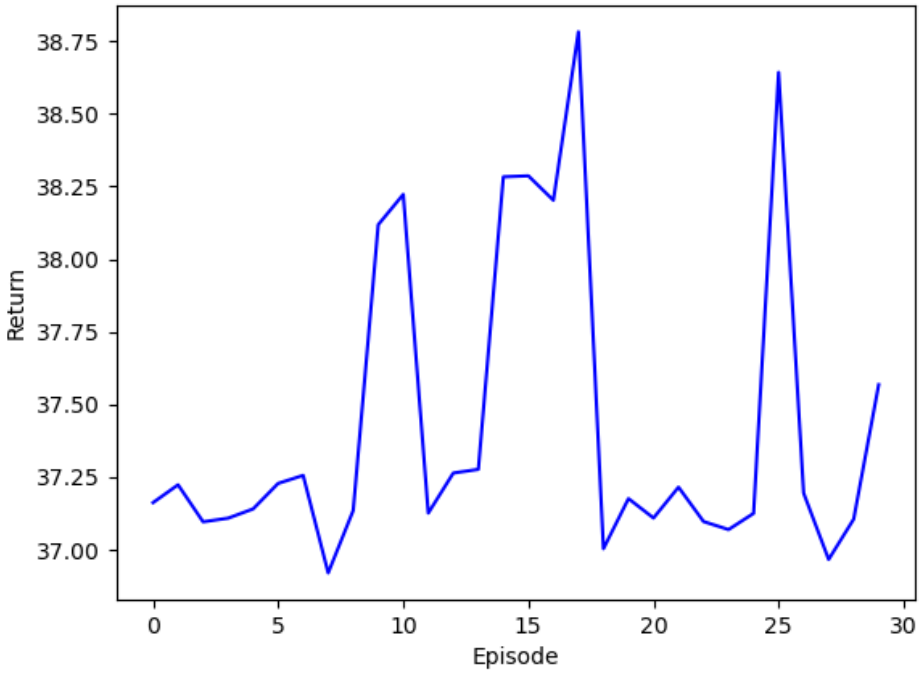

Figura 4: Recompensas da avaliação de política

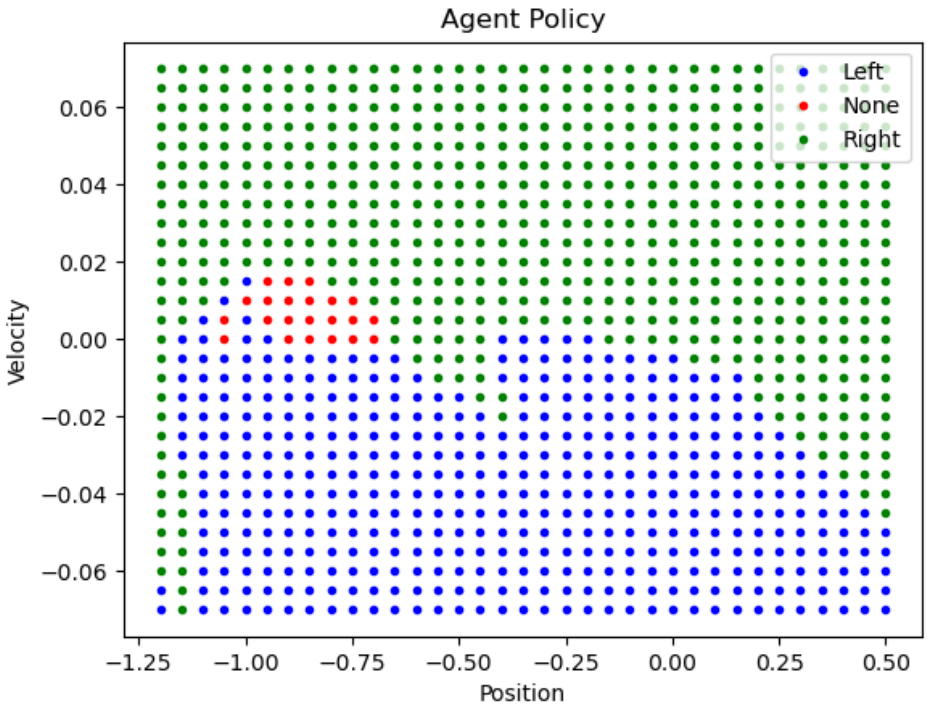

Figura 5: Política adotada

Percebe-se que, após forçar uma boa política, o carrinho conseguiu subir o morro todas as vezes, mesmo com ruídos. Ainda assim, a política encontrada após o treinamento teve resultados suficientemente positivos, provando a eficiência do método. Também é interessante notar o que foi aprendido: O carrinho tende a inicialmente empurrar para frente e depois para trás, tomando um impulso e, ao começar a ir novamente para frente, empurra para a direita para chegar mais longe.