Inteligência Artificial para Robótica Móvel

CT-213

Instituto Tecnológico de Aeronáutica

Relatório do Laboratório 11 - Aprendizado por Reforço Livre de Modelo

Leonardo Peres Dias

21 de junho de 2025

Sumário

1 Breve Explicação em Alto Nível da Implementação

Nesta implementação, utilizamos o algoritmo *Deep Q-Network (DQN)* para resolver o ambiente `MountainCar-v0` da biblioteca *Gymnasium*. O agente é representado pela classe `DQNAgent`, que encapsula os componentes de *Reinforcement Learning* com aproximação de função, como a rede neural, a política de exploração e o buffer de experiência.

A rede neural possui uma arquitetura com duas camadas ocultas de 24 neurônios, utilizando função de ativação ReLU. A camada de saída é linear e estima os valores das ações (Q-values). O treinamento da rede é realizado com a *loss Mean Squared Error (MSE)* e o otimizador *Adam*.

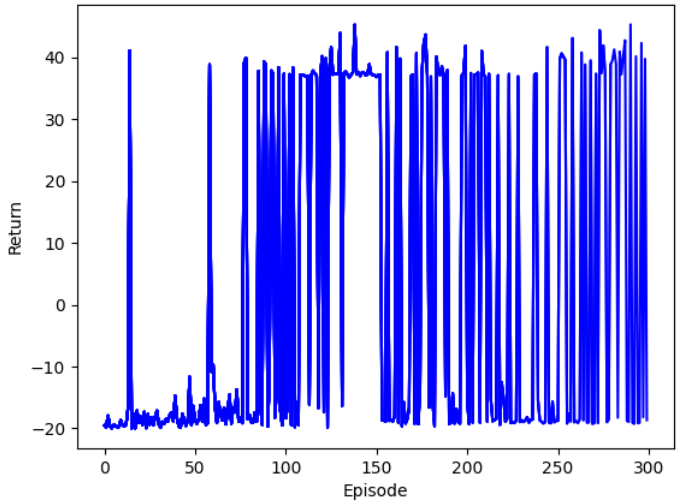
Implementamos também uma técnica de *reward engineering*, com o objetivo de acelerar o aprendizado. A recompensa é aumentada conforme o agente se aproxima do topo da montanha, com um bônus adicional ao alcançar a posição final desejada.

2 Figuras Comprovando Funcionamento do Código

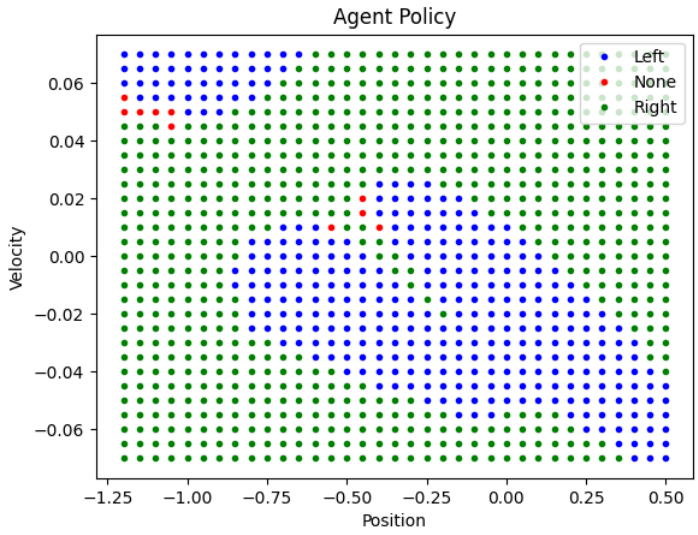
2.1 Sumário do Modelo

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 24)                72

 dense_1 (Dense)             (None, 24)                600

 dense_2 (Dense)             (None, 3)                 75

=================================================================
Total params: 747 (2.92 KB)
Trainable params: 747 (2.92 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

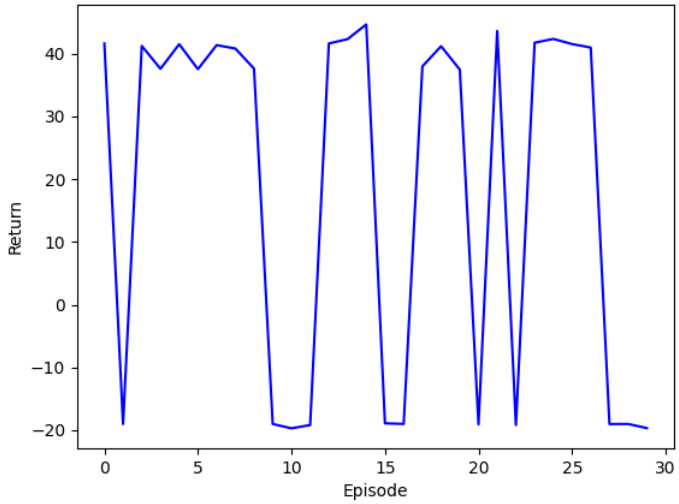
2.2 Retorno ao Longo dos Episódios de Treinamento



2.3 Política Aprendida pelo DQN



2.4 Retorno de 30 Episódios Usando a Rede Neural Treinada



3 Discussão dos Resultados

Durante os 30 episódios de avaliação, o agente obteve desempenho satisfatório em 18 deles, correspondendo a uma taxa de sucesso de 60%. Esses episódios foram caracterizados por pontuações positivas, indicando que o objetivo do ambiente foi alcançado com sucesso. Nos demais 40% dos episódios, o agente falhou em atingir o objetivo dentro do tempo máximo permitido, resultando em pontuações negativas próximas de -19.

O retorno médio ao longo dos episódios foi de aproximadamente 18,80, valor impactado negativamente pelas penalidades acumuladas nos episódios malsucedidos. Observa-se uma contradição entre episódios de alto desempenho e episódios com falha total, o que sugere que a política aprendida ainda apresenta instabilidade. Esse comportamento pode estar relacionado à generalização incompleta da política treinada.

No geral, embora o agente tenha demonstrado capacidade de resolver a tarefa em uma parte significativa dos episódios, um maior treinamento e refinamento das estratégias de aprendizado ainda podem ser usados para melhorar o desempenho do agente.