Reinforcement Learning em Ambientes Gymnasium

Análise e Implementação no Lunar Lander

Pedro Amaro
Rui Almeida

O Ambiente Lunar Lander: O Desafio

O objetivo é uma aterragem segura na plataforma designada, um desafio clássico no Reinforcement Learning.

1

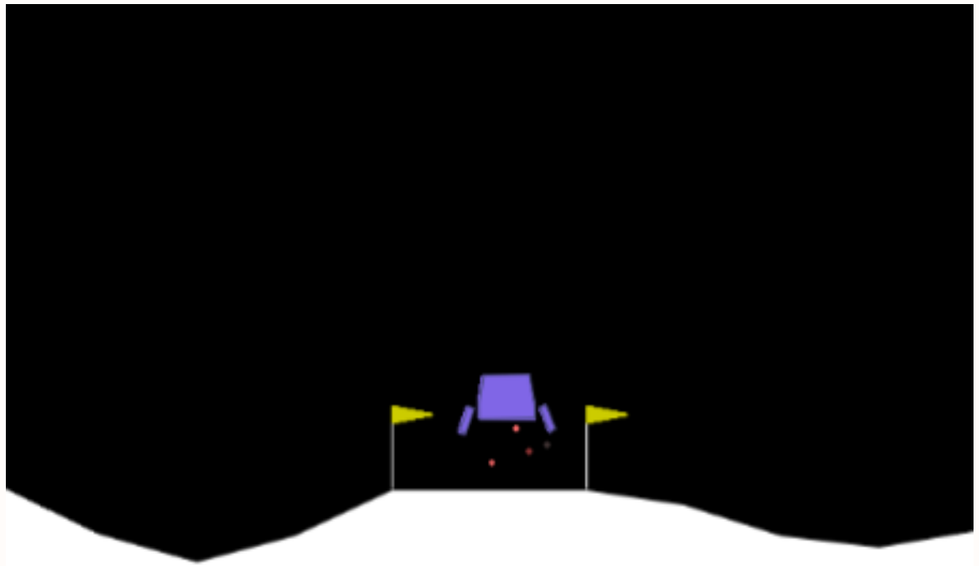
Ações Discretas

Quatro ações distintas: acelerar para cima, inclinar para a esquerda, inclinar para a direita ou permanecer inativo.

2

Espaço de Observação

Oito dimensões de observação que descrevem o estado atual da nave, como posição e velocidade.



3

Limiar de Sucesso

Uma pontuação de 200 ou mais é necessária para considerar a aterragem um sucesso.

Espaço de Observação e Estrutura de Recompensas

O agente aprende através de um sistema de recompensas cuidadosamente calibrado, incentivando aterragens seguras e penalizando falhas.

Espaço de Observação (8 Dimensões)

- **Posição (x, y):** Coordenadas da nave no ambiente.
- **Velocidade (vx, vy):** Velocidade horizontal e vertical da nave.
- **Ângulo (θ) & Velocidade Angular (ω):** Orientação e rotação da nave.
- **Contacto das Pernas:** Indica se a perna esquerda ou direita tocou no solo. (Booleano)

Estrutura de Recompensa

- **Aterragem:** +100 pontos por sucesso.
- **Colisão:** -100 pontos por impacto.
- **Contacto das Pernas:** +10 pontos por cada perna a tocar no solo.
- **Motor Lateral:** -0.03 pontos por cada uso.
- **Motor Principal:** -0.30 pontos por cada uso.

Personalização do Ambiente: Sistema de Combustível

Introduzimos uma restrição de combustível para adicionar complexidade e realismo ao desafio do Lunar Lander.

Capacidade Limitada

O Lander tem uma capacidade de combustível entre 500 e 1000 unidades, tornando a gestão do combustível crucial.

Consumo por Ação

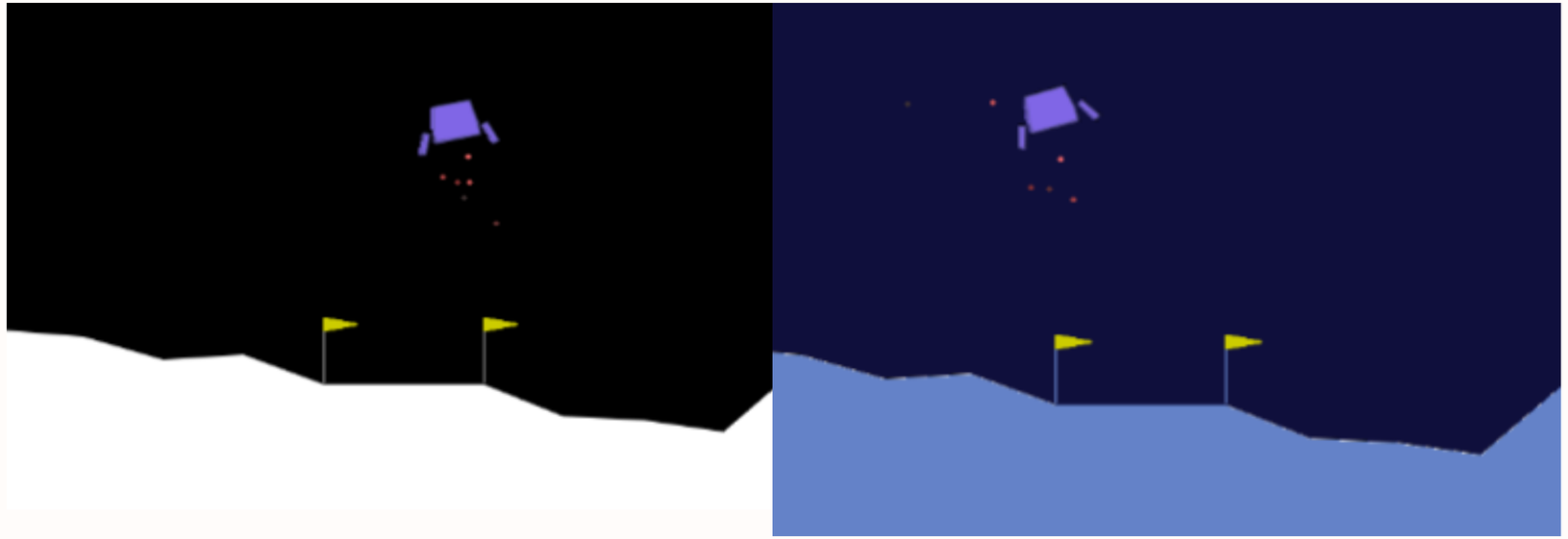
O motor principal consome 5 unidades, enquanto os motores laterais consomem 1 unidade por uso.

Fim do Episódio

Ficar sem combustível resulta na terminação do episódio com uma penalidade de -100 pontos.

Espaço de Observação Expandido

A dimensão do espaço de observação foi aumentada de 8 para 9, incluindo o nível de combustível normalizado.



Personalização do Ambiente: Randomização da Física

Para simular condições mais desafiadoras e imprevisíveis, introduzimos elementos de física estocástica.

Ambiente Original

- **Gravidade Fixa:** Sempre -10.0.
- **Sem Vento:** Ausência total de forças laterais.
- **Sem Turbulência:** Movimento do ar estável.
- **Posição Inicial Fixa:** Ponto de partida inalterado.

Ambiente Personalizado

- **Gravidade Aleatória:** Varia entre -8.0 e -12.0.
- **Força do Vento:** Intensidade entre 5.0 e 20.0.
- **Turbulência:** Fator de 0.5 a 1.5, para maior instabilidade.
- **Posição Inicial Aleatória:** Coordenada X entre 2.0 e 18.0.

Agentes de Reinforcement Learning Testados

Avaliamos dois algoritmos proeminentes de Reinforcement Learning em múltiplas configurações para entender o seu desempenho.

PPO (Proximal Policy Optimization)

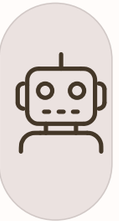
Um algoritmo de gradiente de política "on-policy", conhecido pela sua estabilidade e eficiência. Testado em 2 configurações de hiperparâmetros.

DQN (Deep Q-Network)

Um algoritmo "off-policy" baseado em valor, que utiliza redes neurais para estimar funções Q. Também testado em 2 configurações.

Treinamento Rigoroso

Cada agente foi treinado por 2 milhões de timesteps, seguido de uma avaliação em 100 episódios.

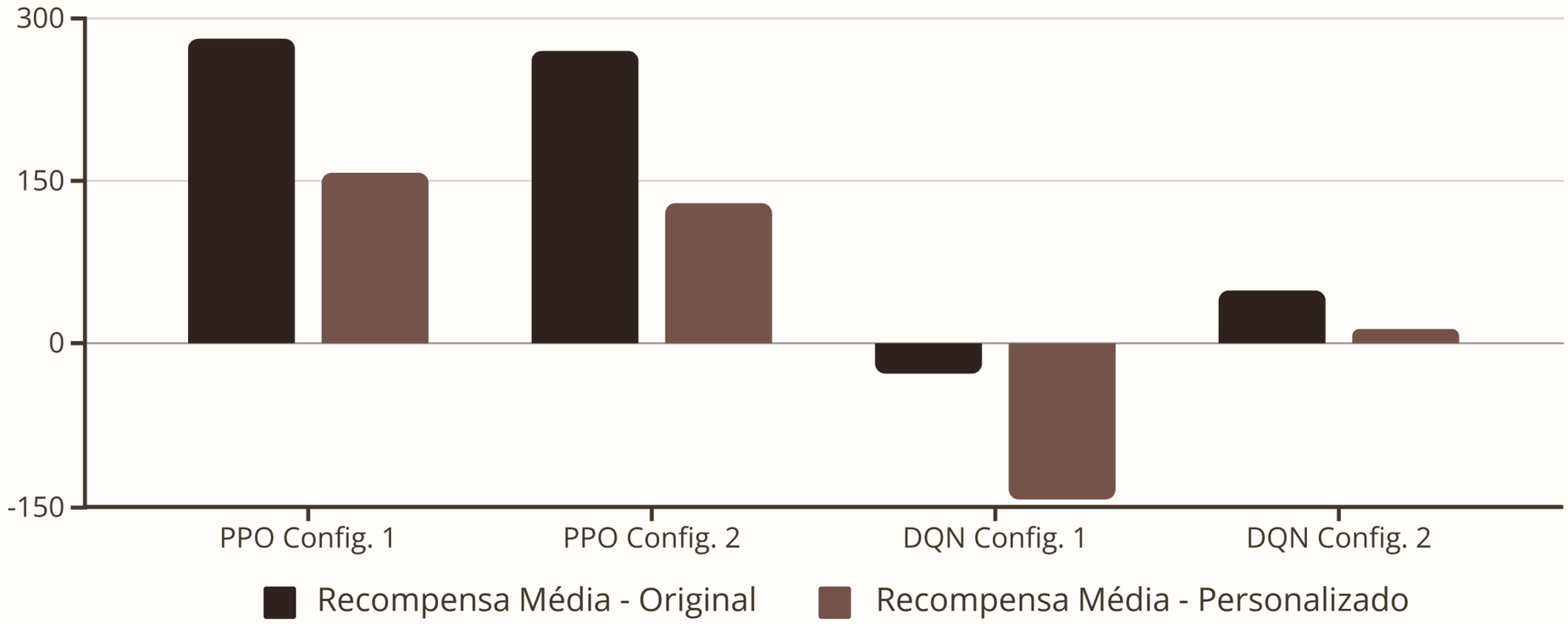


Total de 8 Agentes

No total, 8 agentes foram treinados (2 algoritmos × 2 configurações × 2 ambientes), permitindo uma análise comparativa abrangente.

Resultados: Recompensa Média

A análise da recompensa média revela o desempenho comparativo dos agentes nos ambientes original e personalizado.



Os resultados mostram que o PPO superou o DQN em ambos os ambientes, embora a recompensa média do PPO tenha diminuído no ambiente personalizado devido à complexidade adicionada.

Resultados: Taxa de Sucesso e Duração do Episódio

Além da recompensa, a taxa de sucesso e a duração média do episódio fornecem insights sobre a eficácia e eficiência dos agentes.

Taxa de Sucesso (%)

- **PPO Original S1:** 98% de aterragens bem-sucedidas, demonstrando robustez.
- **PPO Personalizado S1:** 60%, uma redução notável devido à aleatoriedade.
- **DQN Original S1:** 25%, indicando dificuldade em convergir para soluções ótimas.
- **DQN Personalizado S1:** 1%, resultado de colisões frequentes e esgotamento de combustível.

Duração Média do Episódio

- **PPO Original S1:** 214.7 passos, um pouso eficiente e direto.
- **DQN Original S2:** 842.1 passos, comportamento de "pairar" subótimo para evitar penalidades de colisão.
- **PPO Personalizado S1:** 263.8 passos, ligeiramente mais longo devido aos desafios adicionais.
- **DQN Personalizado S1:** 206.1 passos, indicando colisões precoces ou esgotamento de combustível.

Comportamento do Agente em Ação

Uma observação visual do comportamento dos agentes treinados destaca as diferenças nas suas estratégias de aterragem.

1

PPO Original:

Aterragem rápida e decisiva, demonstrando eficácia.

2

PPO Personalizado:

Luta contra as novas condições de vento e gestão de combustível.

3

DQN Original:

Comportamento de pairar subótimo, consumindo mais tempo e combustível.

4

DQN Personalizado:

Colisões frequentes, refletindo a sua incapacidade de se adaptar ao ambiente complexo.

Conclusões e Trabalho Futuro

As nossas descobertas fornecem uma base sólida para futuras investigações em Reinforcement Learning e adaptação de ambientes.

Principais Descobertas

- **PPO vs. DQN:** PPO superou significativamente o DQN em ambos os ambientes.
- **Impacto do Ambiente Personalizado:** O ambiente personalizado reduziu a recompensa PPO em 44% (de 281 para 157).
- **Restrições de Combustível:** Criaram desafios estratégicos, exigindo gestão cuidadosa dos recursos.
- **DQN Subótimo:** O DQN aprendeu um comportamento de pairar com alta duração do episódio (842 passos) e baixa taxa de sucesso (6%).

Trabalho Futuro

- **Modo Queda Livre:** Implementar queda livre após esgotamento de combustível para maior realismo.
- **Otimização Multi-objetivo:** Desenvolver agentes que equilibrem eficiência de combustível e segurança.
- **Transfer Learning:** Explorar a transferência de conhecimento do ambiente original para o personalizado.