Reinforcement Learning with Customised Gymnasium Environments

Assignment 2

Trabalho realizado por:

- Inês Castro up202304060
- Simão Gomes up202304752
- Soraia Costa up202305078

Índice

Ambiente gymnasium: IceHockey-v5

- Jogo de hóquei no gelo para **dois jogadores (agente vs adversário)**
- O adversário é controlado pelo ambiente
- **Objetivo:** Marcar mais golos do que o adversário
- O agente controla **um único jogador**
- O jogo decorre em **tempo contínuo**, com decisões a cada timestep
- A dinâmica do jogo envolve:
  - Movimento constante
  - Disputa pela posse do disco
  - Ataque e defesa simultâneos
- O agente deve aprender:
  - Quando atacar
  - Quando defender
  - Como posicionar-se corretamente

Caracteríticas do Ambiente gymnasium

Percepções

- Frames RGB do ecrã do jogo
- Informação visual parcial, sem acesso ao estado interno do jogo

Estados

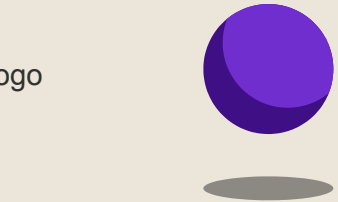
- Construídos através do empilhamento de 4 frames consecutivos
- Permite inferir movimento e direção
- Necessário para capturar a dinâmica temporal do jogo

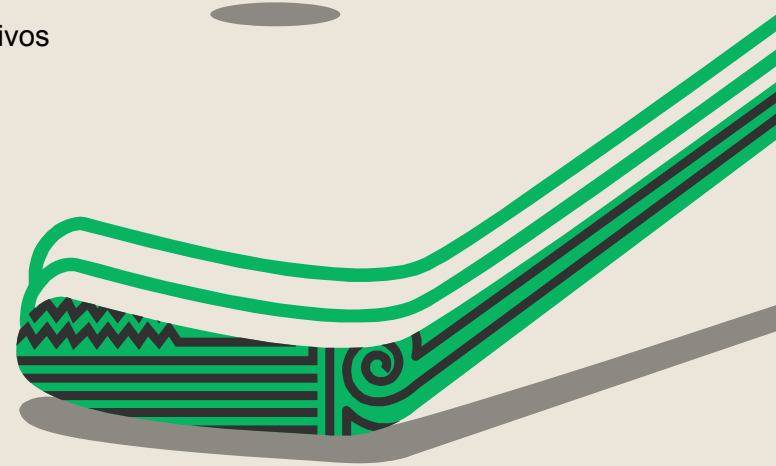
Ações

- Espaço de ações discreto com 18 ações
- Combinações de movimento e ação FIRE
- Ação FIRE essencial para interagir com o disco

Recompensas (ambiente original)

- +1 ao marcar golo
- −1 ao sofrer golo
- Longos períodos sem qualquer feedback





Mudanças introduzidas No ambiente

Propostas Iniciais (Design Conceptual)	Implementação Final (Após Testes Empíricos)
• Manter a **reward objetivo original -** +1 ao marcar golo, −1 ao sofrer golo e Garantir alinhamento com o objetivo real do jogo • Introduzir **orientação direcional -** Recompensar progresso do disco em direção à baliza adversária • Incentivar **posse do disco -** Pequena recompensa por manter controlo • Incentivar **pressão defensiva -** Recompensar redução da distância ao disco • Reforçar **ações taticamente relevantes -** Passes, remates enquadrados e roubos de disco • Penalizar **comportamentos degenerados -** Inatividade prolongada e Proximidade excessiva entre colegas	• **Reward objetivo mantida** +1 ao marcar golo, −1 ao sofrer golo • **Penalização temporal removida** Provocava instabilidade no treino • **Incentivo leve ao movimento** Recompensa para ações de deslocamento • **Incentivo moderado à ação FIRE** Promove interação sem forçar comportamento • **Penalização por repetição excessiva** Evita loops e políticas degeneradas • **Penalização por ausência de remates** Desencoraja passividade prolongada

Algoritmos de Reinforcement Learning

A utilização de múltiplos algoritmos permitiu analisar como diferentes paradigmas de aprendizagem reagem às alterações introduzidas no ambiente.

PPO (on-policy)

conhecido pela sua estabilidade e robustez, especialmente adequado a ambientes complexos e a funções de recompensa modificadas.

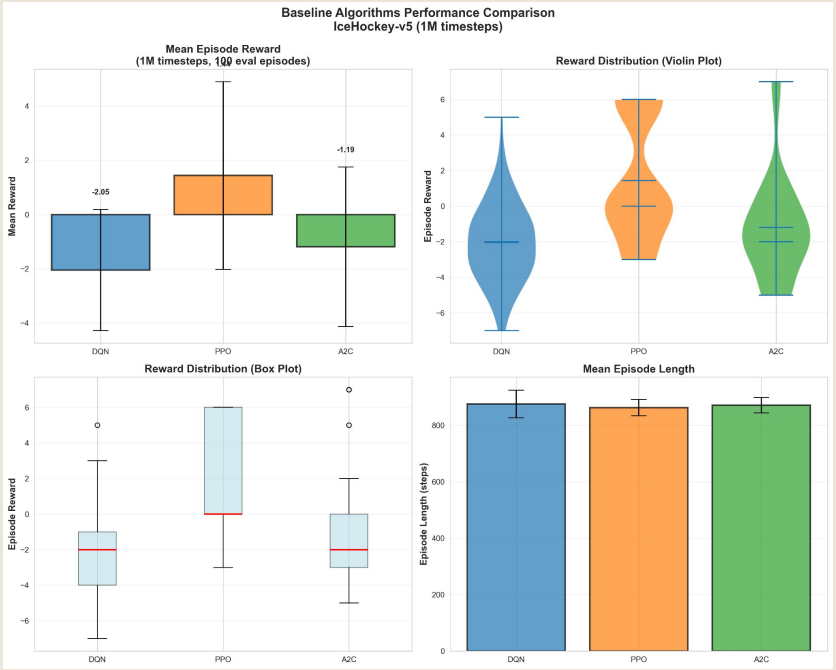
DQN (off-policy)

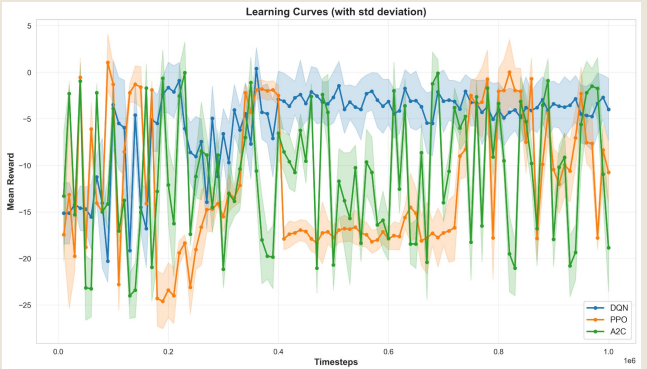
baseado em aprendizagem de valores Q, eficiente em termos de dados, mas mais sensível a alterações na função de recompensa.

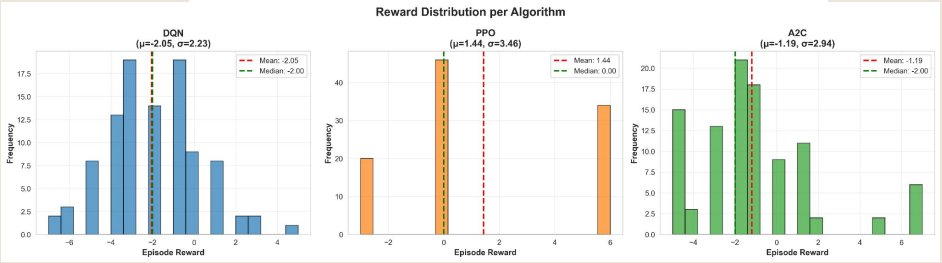
A2C (on-policy)

simples e eficiente no arranque da aprendizagem, mas com menor estabilidade a longo prazo.

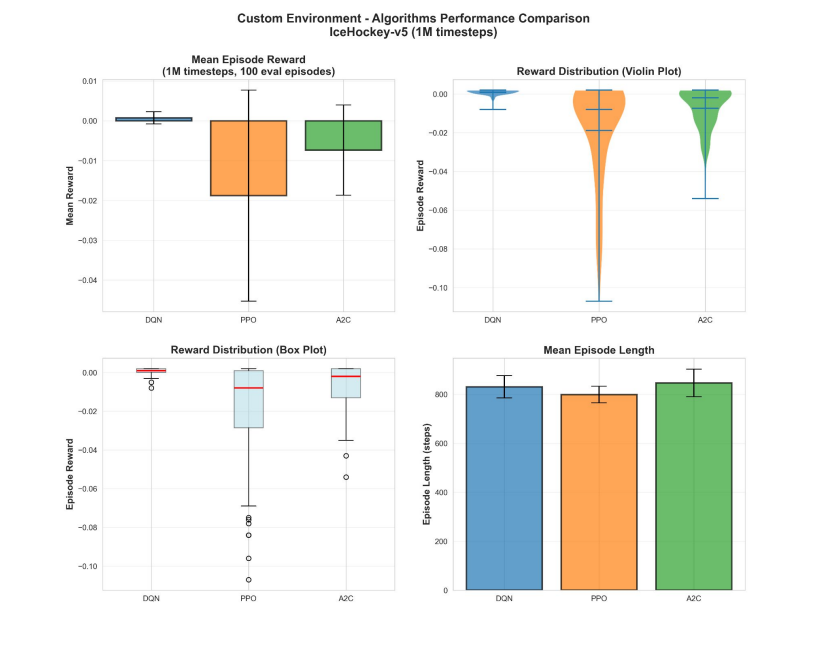
Análise de resultados: Baseline enviroment

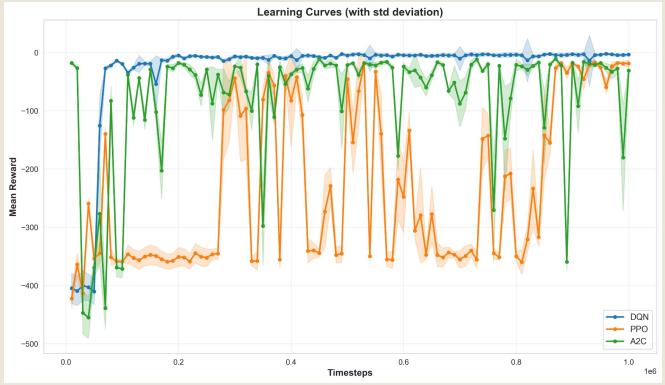


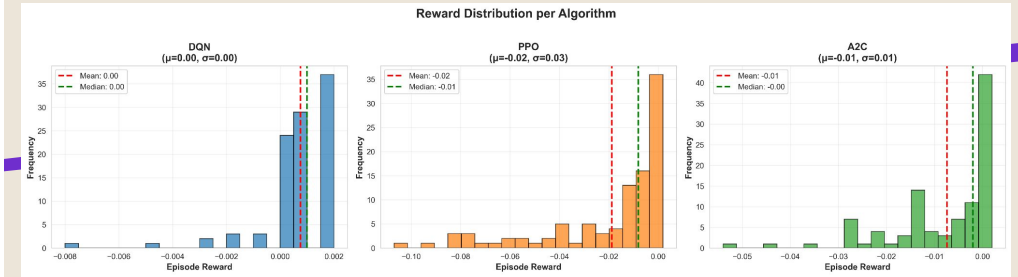




Análise de resultados: custom enviroment

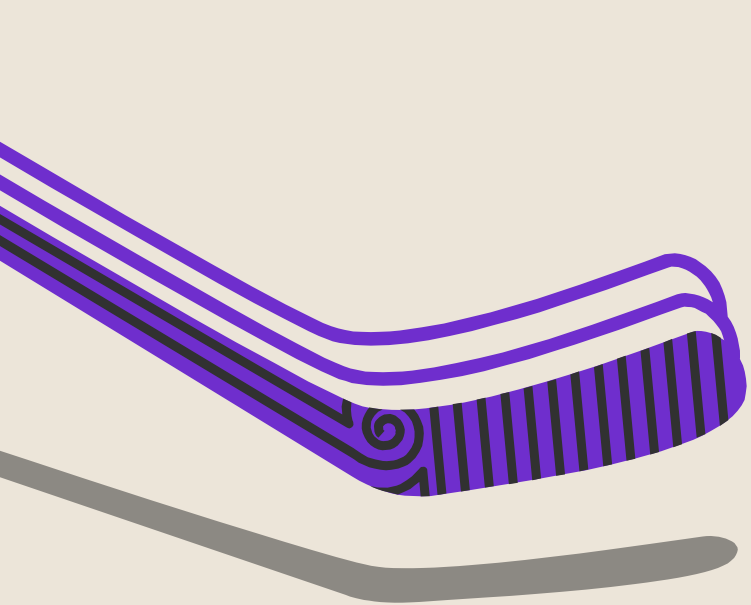






Análise de resultados:

Algoritmo	Baseline (Média ± Desvio)	Custom (Média ± Desvio)	Variação
DQN	−2.05 ± 2.23	+0.001 ± 0.002	↑ Melhoria significativa
PPO	+1.44 ± 3.46	−0.02 ± 0.03	↓ Quebra de desempenho
A2C	−1.19 ± 2.94	−0.01 ± 0.01	↑ Melhoria moderada



03

CONCLUSÃO

Melhoria: Integração de técnicas de **Computer Vision** para extração de características visuais relevantes

Vídeo