Reinforcement Learning Unveiled: Uma ferramenta interativa para a visualização e exploração do aprendizado por reforço com o algoritmo Q-learning

Gustavo tironi
FGV EMAp
Fundação Getulio Vargas (FGV)
Rio de Janeiro, Brasil
gustavo.tironi@fgv.edu.br

Kauan Mariani Ferreira
FGV EMAp
Fundação Getulio Vargas (FGV)
Rio de Janeiro, Brasil
b51058@fgv.edu.br

Pedro Henrique Coterli
FGV EMAp
Fundação Getulio Vargas (FGV)
Rio de Janeiro, Brasil
b51063@fgv.edu.br

***Resumo*—O Aprendizado por Reforço (RL) é uma ramificação que faz parte dos grandes pilares da "Inteligência Artificial". Apesar de muito relevante e altamente aplicado, os seus conceitos fundamentais e a dinâmica de seus algoritmos podem ser difíceis de compreender. Por conta disso, desenvolvemos como trabalho final da disciplina de Visualização de Dados o "Reinforcement Learning Unveiled", um sistema de visualização interativo, projetado para desmistificar o aprendizado por reforço, com foco no algoritmo Q-learning. A ferramenta apresenta um agente, estilizado como Pac-Man, navegando em um ambiente de grade (Grid World) com o objetivo de alcançar a cereja enquanto evita obstáculos, os fantasmas. Os usuários podem configurar dinamicamente os parâmetros do ambiente e do algoritmo de RL, como a taxa de aprendizado (α), o fator de desconto (γ) e a estratégia de exploração (ϵ). O sistema oferece um painel com múltiplas visualizações coordenadas que mostram o processo de aprendizado, incluindo a taxa de sucesso ao longo do tempo, a política aprendida, um mapa de calor dos valores-Q e gráficos detalhados da evolução dos valores-Q para células específicas. O sistema proporciona uma experiência de aprendizado imersiva que agrega a teoria e a prática, permitindo que os usuários observem diretamente como suas escolhas de parâmetros impactam o comportamento e o desempenho do agente.**

I. Introdução

O Aprendizado por Reforço (RL) [5] estabeleceu-se como um paradigma poderoso para treinar agentes autônomos a tomar decisões ótimas em ambientes complexos, com aplicações notáveis em robótica, jogos e finanças. Diferentemente do aprendizado supervisionado, o RL não depende de um conjunto de dados rotulado. Em vez disso, um agente aprende através da interação direta com um ambiente, recebendo feedback na forma de recompensas ou penalidades. Embora eficaz, esse paradigma de aprendizado por experimentação pode representar uma barreira de entrada significativa para estudantes e entusiastas, que muitas vezes têm dificuldade de entender como a iteração leva ao aprendizado.

Essa dificuldade reside em visualizar os conceitos abstratos que governam o comportamento do agente, como a "política"(a estratégia do agente) e a "função de valor"(a desejabilidade de estar em um determinado estado). Observar essas estruturas evoluindo ao longo do tempo é crucial para uma compreensão profunda.

Este trabalho apresenta o *Reinforcement Learning Unveiled*, uma ferramenta de visualização interativa projetada para preencher essa lacuna. Nosso objetivo é fornecer uma plataforma hands-on onde os usuários possam não apenas aprender e observar um algoritmo de RL em ação, mas também manipular seus componentes fundamentais e ver as consequências em tempo real. O sistema foca no algoritmo Q-learning, implementado em um cenário clássico de Grid World, tornando o processo de aprendizado tangível e acessível.

II. Métodos

A. Aprendizado por Reforço

O framework de RL é formalizado como um Processo de Decisão de Markov (MDP) [5], que consiste nos seguintes componentes:

- **Agente**: A entidade que aprende e toma decisões. Em nosso sistema, é o Pac-Man.
- **Ambiente**: O mundo externo com o qual o agente interage. No nosso caso, um Grid World.
- **Estado** (s): Uma representação da situação atual do agente no ambiente. Aqui, a posição (x, y) do Pac-Man na grade.
- **Ação** (a): Uma das possíveis ações que o agente pode executar. Para o Pac-Man, as ações são mover para Cima, Baixo, Esquerda ou Direita.
- **Recompensa** (r): Um sinal numérico que o ambiente fornece ao agente após uma ação. O objetivo do agente é maximizar a recompensa cumulativa.
- **Política** (π): A estratégia do agente, que mapeia estados a ações. A política define o comportamento do agente em um determinado estado.

O agente e o ambiente interagem em um ciclo: o agente observa um estado s, executa uma ação a, recebe uma re-

compensa r e transita para um novo estado s'. Este processo se repete até que um estado terminal seja alcançado.

B. Q-Learning

Q-learning [6] é um algoritmo de RL model-free (não requer um modelo do ambiente) e off-policy (pode aprender a política ótima independentemente das ações tomadas). Seu objetivo é aprender uma função de valor-ação, $Q(s, a)$, que estima a recompensa total esperada ao executar a ação a no estado s e seguir a política ótima a partir de então.

Os valores-Q são geralmente armazenados em uma tabela chamada Q-table e são atualizados iterativamente usando a equação de Bellman:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad (1)$$

Onde:

- α é a **taxa de aprendizado**, que determina o peso dado à nova informação.
- γ é o **fator de desconto**, que valoriza as recompensas futuras.
- $\max_{a'} Q(s', a')$ é a estimativa da recompensa futura ótima a partir do próximo estado s'.

Um desafio central em RL é o trade-off entre **exploração** (tentar novas ações para descobrir suas recompensas) e **explotação** (usar o conhecimento atual para maximizar a recompensa) [3]. O Q-learning frequentemente lida com isso através de uma política ϵ-greedy, onde o agente escolhe uma ação aleatória com probabilidade ϵ e a melhor ação conhecida com probabilidade $1 - \epsilon$. Tipicamente, ϵ decai ao longo do tempo (Epsilon Decay) para favorecer a explotação à medida que o agente ganha mais experiência [5].

III. Trabalhos Relacionados

A ideia e realização do nosso projeto foi influenciada por diversas ferramentas e publicações existentes que buscam desmistificar conceitos complexos de aprendizado de máquina por meio da interatividade e da visualização.

A ideia foi diretamente inspirada pelo projeto *Transformer Explainer* [1]. Embora aborde um tópico totalmente distinto (arquiteturas Transformer), a ideia de decompor um sistema complexo em componentes visuais e interativos para facilitar a compreensão foi uma inspiração fundamental para a nossa ferramenta. Da mesma forma, a publicação na plataforma Distill "Visualizing and Understanding Reinforcement Learning" [2] serviu de inspiração para a forma de mostrar e de construir a nossa visualização, já que também aborda um tópico de RL usando visualizações didáticas.

Em relação à explicação teórica e apresentação dos conceitos, o curso "Deep Reinforcement Learning Course"da Hugging Face [4] foi uma referência importante. A forma como o curso guia os alunos desde os fundamentos teóricos até a implementação prática influenciou a nossa decisão de estruturar o sistema em páginas progressivas, que vão desde uma introdução simplificada, definições formais até o ambiente iterativo final.

IV. Resultados

O sistema final é uma aplicação web de múltiplas páginas desenvolvida com o framework Svelte. A estrutura do sistema foi pensada para guiar o usuário em uma jornada de aprendizado progressiva:

1) **Introdução Intuitiva:** Uma página inicial que explica o "o quê"da RL de forma lúdica.
2) **Definição Formal:** Uma segunda página aprofunda os conceitos, definindo formalmente Agente, Ambiente, Recompensas e Política.
3) **Explicação do Q-Learning:** Uma página dedicada a explicar como o agente aprende. Ela apresenta a equação de Bellman de forma decomposta e inclui uma visualização interativa de uma Q-Table sendo atualizada passo a passo (Fig. 1).
4) **Visualização Interativa Principal:** O coração do sistema, onde o usuário pode experimentar livremente. É uma página "sandbox"onde todos os conceitos se unem.

Essa abordagem segmentada permite que cada usuário engaje com o conteúdo no nível de profundidade que for mais adequado e o guia para se aprofundar até chegar na visualização final.

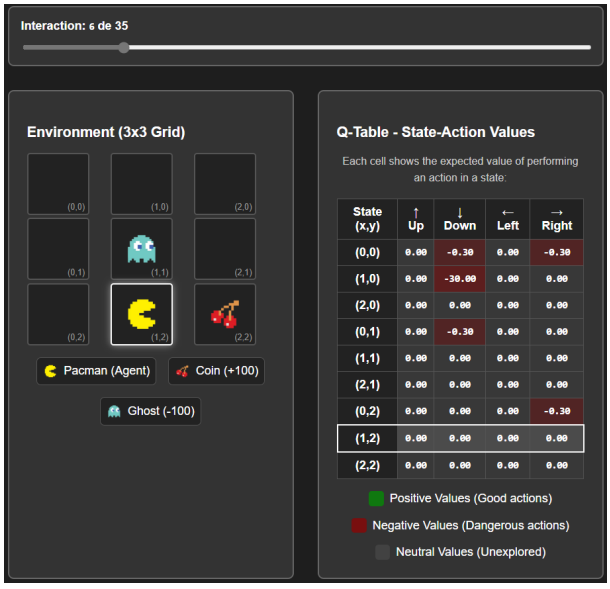

Figura 1. Explicação interativa da atualização da Q-Table.

A. Visualização Interativa Principal

Esta página (Fig. 2) é onde o aprendizado ativo acontece. Ela é composta por um painel de controle e múltiplos painéis de

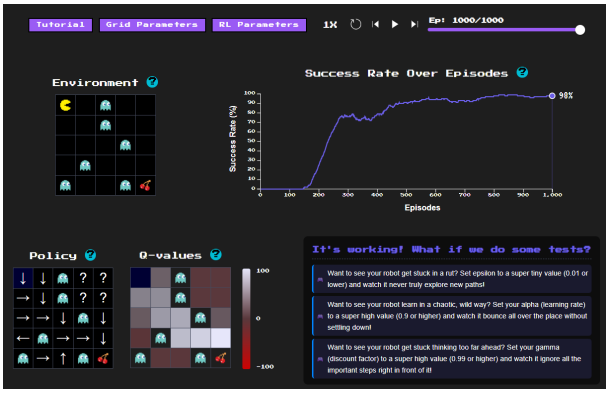

Figura 2. Visualização interativa do aprendizado com Q-learning.

visualização coordenados, que são atualizados após a execução do algoritmo de Q-Learning diretamente no navegador.

O usuário tem total controle sobre o experimento através de duas abas:

- **Grid Parameters:** Permite modificar a largura e a altura do ambiente. O usuário pode clicar nas células do grid para adicionar ou remover obstáculos (fantasmas), personalizando o desafio para o agente.
- **RL Parameters:** Oferece sliders para ajustar os hiperparâmetros do algoritmo: Alfa, Gama, Épsilon, Decaimento de Épsilon, Número de Episódios e Máximo de Passos por Episódio.

Após clicar em ”Apply”, o algoritmo Q-Learning é executado com os parâmetros definidos, e os resultados (histórico de recompensas, Q-Table final, etc.) são armazenados para alimentar as visualizações.

1) Visão Geral do Treinamento: Localizada na parte superior da interface, esta seção oferece um resumo do desempenho geral do agente.

- **Environment View**: Uma representação visual do Grid World, mostrando a posição do agente, dos fantasmas e do objetivo (cerejas).
- **Success Rate Over Episodes**: Um gráfico de linha que exibe a taxa de sucesso (episódios em que o objetivo foi alcançado) ao longo do tempo. Esta visualização é fundamental para entender padrões de aprendizado de maneira mais ampla.

2) Visualização da Política e dos Valores Globais: Quando nenhuma célula específica da grade está selecionada, a parte inferior da interface exibe informações globais sobre a solução aprendida.

- **Policy View**: Mostra a política ótima aprendida. Cada célula exibe uma seta indicando a melhor ação a ser tomada naquele estado, correspondendo a $\arg\max_a Q(s,a)$. Células com um ponto de interrogação (”?”) representam estados que não foram explorados ou que estão com ações ”empatadas”, sem a definição de uma política ótima absoluta.
- **Q-values Heatmap**: Um mapa de calor que visualiza o valor de cada estado, definido como $V(s) = \max_a Q(s,a)$. A cor de cada célula representa a recompensa máxima esperada a partir daquele estado.

3) Análise Detalhada por Célula: Ao clicar em qualquer célula no *Environment View* principal, as visualizações inferiores mudam (Fig. 3) para fornecer uma análise detalhada daquele estado específico.

- **Q-Values for Cell**: Uma visualização que exibe os quatro valores-Q (para as ações Cima, Baixo, Esquerda, Direita) do estado selecionado. A cor de cada seta corresponde ao valor-Q daquela ação no episódio atualmente em exibição.
- **Q-Values Over Episodes for Cell**: Um gráfico de linha que plota a evolução dos quatro valores-Q para o estado selecionado ao longo de todos os episódios de treinamento.

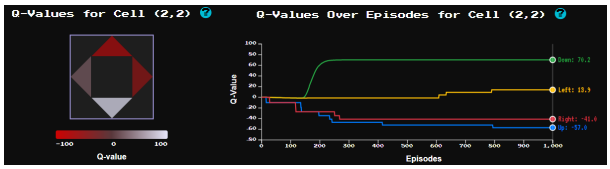

Figura 3. Visualizações para uma célula selecionada

V. Discussão

O projeto demonstra como a visualização interativa pode servir como uma poderosa ferramenta pedagógica. O principal resultado ao usuário que a nossa ferramenta proporciona é a desmistificação do processo de aprendizado em RL, tornando-o tangível e menos abstrato. Ele permite que os usuários construam uma intuição sobre o impacto dos hiperparâmetros, uma habilidade crucial em RL. Ao interagir com o sistema, os usuários rapidamente estabelecem uma conexão direta entre suas ações no painel de controle e o comportamento emergente do agente.

Por exemplo, a visualização mostra de forma clara o dilema exploração-explotação. Um usuário que define um valor de épsilon (ϵ) muito baixo (ex: 0.01) observa o agente ficar ”preso em uma rotina”, seguindo um caminho subótimo sem nunca explorar alternativas que poderiam ser mais eficientes. Em contrapartida, ao definir uma taxa de aprendizado (α) muito alta (ex: 0.9), o usuário testemunha um aprendizado caótico. Similar, ao manipular o fator de desconto (γ), o usuário aprende sobre a ”visão de futuro”do agente. Um valor de γ muito alto (ex: 0.99) pode fazer o agente ponderar tanto as recompensas distantes que ele se torna lento para aprender os passos críticos mais iniciais.

Ao experimentar com essas configurações por um tempo, o usuário começa a entender como cada parâmetro afeta o

aprendizado de maneira única. Essa compreensão prática é fundamental, preparando o usuário para aplicar esses e outros algoritmos de aprendizado de máquina de forma mais eficaz no futuro.

VI. Trabalhos Futuros

Existem diversas adições possíveis ao nosso projeto que podem aprofundar ainda mais a experiência de aprendizado. Essas adições são:

- **Inclusão de Novos Algoritmos de Aprendizagem**: O sistema poderia ser expandido para incorporar outros algoritmos clássicos de RL.
- **Comparação Direta entre Algoritmos**: Uma funcionalidade relevante seria a capacidade de executar diferentes algoritmos lado a lado, no mesmo ambiente. Isso permitiria que os usuários comparassem visualmente, episódio por episódio, as diferenças na velocidade de convergência, na política final e as diferenças no aprendizado dos métodos.
- **Ambientes de Aprendizagem Diversificados**: Para além do Grid World, poderíamos incluir novos tipos de ambientes, com dinâmicas que não se baseiam em grades.
- **Introdução de Ambientes Estocásticos**: Uma melhoria possivelmente interessante seria a adição de ambientes não-determinísticos (estocásticos), onde uma ação pode levar a diferentes estados com certas probabilidades. Isso introduziria o conceito de incerteza.

VII. Conclusão

Este artigo apresentou o *Reinforcement Learning Unveiled*, um sistema de visualização interativo projetado para tornar os conceitos de Aprendizado por Reforço, especificamente o Q-learning, mais acessíveis e compreensíveis. Ao combinar um ambiente de Grid World personalizável com um painel de visualizações coordenadas e interativas, a ferramenta permite que os usuários explorem ativamente a relação entre os hiperparâmetros do algoritmo e o comportamento emergente do agente. Além disso, a plataforma conta com guias teóricos que apresentam, de forma formal e acessível, os fundamentos do aprendizado por reforço e do algoritmo Q-Learning, acompanhando o usuário até que ele esteja preparado para explorar a visualização interativa. O projeto serve como um exemplo prático do potencial da visualização para desmistificar algoritmos de inteligência artificial, fomentando uma compreensão mais profunda e intuitiva para uma nova geração de estudantes e profissionais da área.

Referências

[1] Jay Alammar. The illustrated transformer. https://jalammar.github.io/illustrated-transformer/, 2018. Acessado em: 15 de junho de 2025.

[2] Dumitru Erhan, Ryan Lange, garnished, et al. Visualizing and understanding reinforcement learning. *Distill*, 2020.

[3] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[4] Thomas Simonini and Hugging Face. Deep reinforcement learning course. https://huggingface.co/learn/deep-rl-course/unit0/introduction, 2022. Acessado em: 15 de junho de 2025.

[5] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. Adaptive computation and machine learning. MIT Press, 2018.

[6] Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. Phd thesis, University of Cambridge, Cambridge, England, 1989.