Análise da variação de diferentes redes neurais e modelos de reward para o problema do Cart Pole

Luis Rigon
CT-213
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brasil
luis.rigon.8771@ga.ita.br

Rafael Louvain
CT-213
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brasil
rafael.louvain.8777@ga.ita.br

Raul Silveira
CT-213
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brasil
raul.silveira.8778@ga.ita.br

I. Introdução Teórica

Vital na gama das inteligências artificiais, o aprendizado de máquina é um dos setores mais estudados no mundo da robótica atualmente. Diversos métodos distintos já foram desenvolvidos e continuam sendo foco de estudos daqueles que buscam compreender mais sobre o uso de inteligência artificial na robótica.

O aprendizado por reforço mostra-se como uma grande alternativa para atingir os objetivos desejados sem partir de dados anotados, como no aprendizado supervisionado.

Foi estudado o uso do *Deep Q Learning*, que vale de uma rede neural para a aproximação da tabela ação-valor, a fim de que fosse automatizado e acelerado o treinamento. A partir de uma quantidade limitada de estados e ações, escolhe-se uma política e executa-se uma nova ação. O agente também utiliza um histórico de episódios a fim de melhorar seu treinamento. O agente então é recompensado, de forma a escolher as ações com melhores recompensas ao longo do treino e assim realizar as melhores ações na avaliação do aprendizado.

II. Objetivo

O *Deep Q Learning* pressupõe duas partes substancialmente relevantes: a implementação do padrão de recompensa e a escolha da rede neural adequada. Nesse sentido, o objetivo do Projeto se concentra no estudo de diferentes combinações de esquemas de recompensa com rede neurais a fim analisar qual das combinações produz melhor resultado para a resolução do problema do *Cart Pole*, bem como adquirir dados relevantes individualmente sobre as arquiteturas das redes neurais e os esquemas de recompensa empregados.

III. Problema e Metodologia

O problema do *Cart Pole* consiste em uma simulação de uma barra verticalmente posicionada sobre um carrinho com liberdade de movimento no eixo x, de forma que a simulação é encerrada caso a posição em x do carrinho diste número maior ou igual a 2.4 unidades e o ângulo entre a barra e a vertical seja maior ou igual a 12 °. Portanto, a resolução do problema é pautada por dois relevantes parâmetros: o *score*, que consiste em uma normalização do total de recompensas acumuladas ao longo dos passos de tempo em uma iteração e do *time*, que considera o tempo que o algoritmo conseguiu permanecer sem cair em alguma das condições de encerramento, sendo o tempo máximo do episódio de 200. Assim, será utilizado o algoritmo do *Deep Q-Networks* (DQN), que será melhor detalhado no próximo tópico, como estrutura principal para a resolução do problema do *Cart Pole*. Assim, considerando as etapas de treinamento e avaliação do agente presentes a cada vez que é executado o DQN, o método utilizado a fim de atingir o objetivo de realizar uma comparação satisfatória entre as diferentes combinações de redes neurais e recompensas consiste em aplicar, para cada combinação, uma única vez o treinamento e partir para a avaliação do agente. Por vezes, o agente ainda não conseguiu fazer convergir o comportamento para o valor desejável de *score* e de tempo com apenas um treinamento, porém, a fim de avaliação de qual algoritmo desempenha melhor deve-se aplicar o mesmo procedimento e portanto, analisa-se o desempenho das combinações todas submetidas a apenas um treinamento. Para isso, é necessário, a cada finalização da avaliação de uma dada combinação,que seja apagado o arquivo cartpole.h5, que armazena o histórico do treinamento, dado que não é desejável que um novo treinamento utilize dos dados do treino anterior.

IV. Implementação

A resolução do problema via *Deep Q-Networks* pressupõe a criação do *environment* do *Cart Pole*, a criação de um modelo de *Deep Learning* e a criação de um agente com uma política relacionada. Nesse sentido, o *environment* foi importado da plataforma da Open Gym e as demais estuturas necessárias ao objetivo do Projeto teve o arcabouço estrutural do Laboratório 12 para a sua implementação. Desse modo, os principais tópicos desenvolvidos na implementação do código estão abaixado elencados.

A. Redes neurais

Para avalição das diferentes redes neurais, foram utilizadas duas arquiteturas, com diferentes funções e número de camadas e neurônios. A primeira arquitetura consiste de uma rede neural bem simples, com apenas duas camadas ao passo que a segunda estrutura foi sugerida por uma referência da literatura[1] como alternativa para a solução do problema do *Cart Pole*. Assim, a primeira rede neural surgiu justamente para fazer alternativa à segunda rede neural que possui quatro

camadas com número de neurônios bem superior e relativamente mais complexa. Desse modo, pretende-se comparar a arquitetura apontada na literatura com a estrutura mais simples e com menos neurônios.

TABLE I
PRIMEIRA ARQUITETURA DA REDE NEURAL

Camada	Neurônios	Função de ativação
Dense	28	tanh
Dense	`action size`	linear

TABLE II
SEGUNDA ARQUITETURA DA REDE NEURAL

Camada	Neurônios	Função de ativação
Dense	512	ReLU
Dense	256	ReLU
Dense	128	ReLU
Dense	`action size`	linear

B. Política

Foi ultilizada a política ϵ-*greedy*, sendo ϵ um hiperparâmetro definido de acordo com a heurística para cada problema (no caso, foi adotado o valor de 0.5), Assim, um número randômico é tomado entre 0 e 1, sendo esse número maior que o valor de ϵ, é retornada a ação gulosa. Caso contrário, é retornada ação aleatória.

C. Recompensa

Inerente ao *environment* do *Cart Pole*, há uma recompensa padrão atrelada. Essa recompensa consiste de um acréscimo +1 para cada estado que respeite as limitações espaciais do *environment* já abordadas.

Existe, também, um modelo artificial de recompensa. A motivação desse tipo de recompensa está relacionada a possibilidade de fornecer uma orientação para o aprendizado de reforço ao longo do processo para algoritmos que só possuem recompensa padrão ao final de muitas etapas além de ser uma opção para reforçar e priorizar estados em detrimentos de outros, atribuindo uma recompensa maior para um dado comportamento preferível.Nesse sentido, quatro padrões diferentes de recompensa foram implementados, sendo abaixo detalhados

O primeiro deles contempla apenas a recompensa padrão do *environment*.

O primeiro modelo artificial proposto, aqui denotado por "Recompensa 1", buscava recompensar baixas velocidades, proximidade com a posição inicial, pequenos ângulos na barra e menores velocidades angulares. Desta forma todos estes fatores recebiam recompensa da seguinte forma, sendo que o valor máximo é tamponado em 1:

$$reward = \frac{10^{-3}}{|Fator|} \tag{1}$$

Além disso, este modelo recompensava a duração do processo, com uma recompensa adcional no valor de um centésimo do marcador temporal.

O próximo padrão de recompensa, denotado por "Recompensa 2", foi fornecido em uma fonte da literatura [2] como um modelo eficiente de recompensa artifical para o problema do *Cart Pole*. A equação que rege o esquema de recompensa está abaixo elencada.

$$reward = 1 - \frac{x^2}{11.52} - \frac{\theta^2}{288} \tag{2}$$

Além disso, na mesma fonte foi apontado que um bom modelo de recompensa artifical contemplava função que tinha o seu valor zerado para as condições indesejadas do problema. Nesse sentido, visando aplicar e testar essa ideia proposta na fonte, o grupo propôs mais um modelo de recompensa, denotado por "Recompensa 3". Esse modelo têm como princípio justamente que a recompensa artificial seja nula no momento em que o estado atinge alguma das condições limites e varia de forma contínua a um valor cada vez menor conforme a simulação se aproxima desse estado. Além disso, considera-se o produto dos dois graus de liberdade descritos no problema, sendo feita uma normalização para que o maior valor possível de recompensa seja 1 e não contraste com o valor da recompensa padrão. Assim, o esquema de recompensa 3 possui sua equação descrita pela fórmula abaixo.

$$reward = \frac{|(2.4 - x) * (0.2094395 - \theta)|}{2.4 * 0.2094395} \tag{3}$$

Sendo 2.4 a máxima posição em x permitida em módulo e 0.2094395 o maior desvio angular permitido. Considera-se, assim, um produto das variações espacial e angular normalizado.

V. RESULTADOS E DISCUSSÃO

A seguir, veremos os resultados obtidos para o problema do CartPole. Em cada tópico é discriminado o resultado relativo ao esquema de recompensa aplicado, sendo cada uma das recompensas submetidas ao treinamento da rede neural e avaliação da política com as duas redes neurais.

A. Recompensa Padrão

O ambiente do Cart Pole, que fora importado da plataforma da Open Gym, pressupõe uma recompensa padrão de 1 para cada estado que respeite as limitações impostas da posição x do carro ser inferior, em módulo, a 2.4 e o ângulo da barra com a vertical ser inferior a 12 °. Desse modo, foi aplicada apenas essa recompensa nas duas redes neurais para a resolução do problema do Cart Pole a fim de comparação com os próximos esquemas de recompensa implementados. Abaixo, constam os resultados obtidos com a execução do algoritmo.

1) Primeira Rede Neural:

- Treinamento

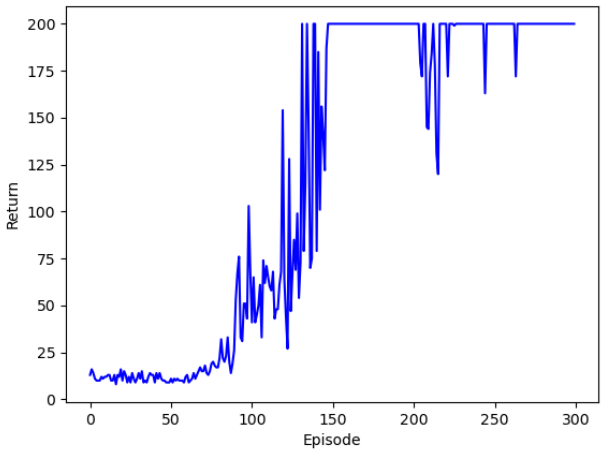

Fig. 1. Recompensa ganha pelo agente por episódio de treinamento.

- Avaliação

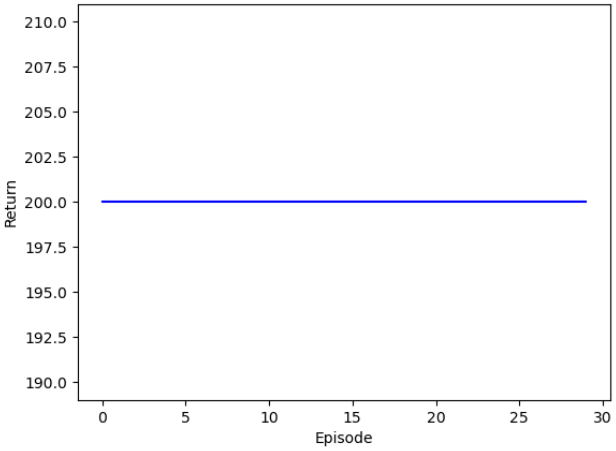

Fig. 2. Recompensa ganha pelo agente por episódio no processo de avaliação.

- Retorno
  O retorno médio da avaliação foi 200, ou seja, para todos casos de avaliação, o agente conseguiu manter o poste de pé em cima do carro, baseado no único treinamento mostrado anteriormente.

Percebe-se que o treinamento do agente foi muito bem sucessido, visto que ele era recompensado somente pelo tempo que mantinha o poste em pé, aliado a uma rede neural eficiente e simples. É notável que a avaliação foi perfeita, mesmo só com um treinamento. Contundo, notou-se que o carrinho assumia um comportamento com tendencia de ir aos extremos do ambiente, fugindo do centro. Dessa forma, nas próximas rewards foi buscado melhorar o comportamento do agente.

2) Segunda Rede Neural:

- Treinamento

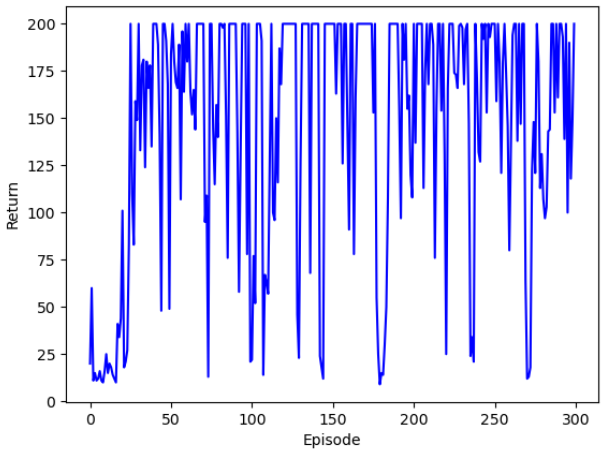

Fig. 3. Resultado do primeiro treinamento.

- Avaliação

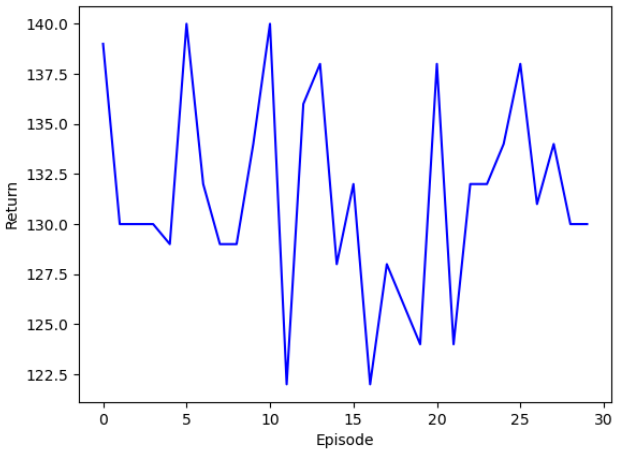

Fig. 4. Retorno por episódio na avaliação.

- Retorno
  O retorno médio foi de 131.366, não alcançando o valor esperado.

No caso da segunda rede, pode-se perceber um treinamento muito inconsistente, atingindo o valor máximo e logo após tendo um episódio curto. O que mostra que o agente não conseguiu aprender perfeitamente com um só treinamento. Comprova-se isso ao analisar a avaliação da rede, não houve nenhum valor ótimo, visto que o agente não terminou seu treinamento, dessa forma, normalizou seus resultados por volta de 130 de score.

B. Recompensa 1

Esse esquema de recompensa implementado pelo grupo considera variações inferiores a 10^{-3} como parâmetro para a comparação. Dessa forma, recompensa o agente por tentar

manter tanto uma posição central quanto velocidades pequenas.

1) Primeira Rede Neural:

- Treinamento

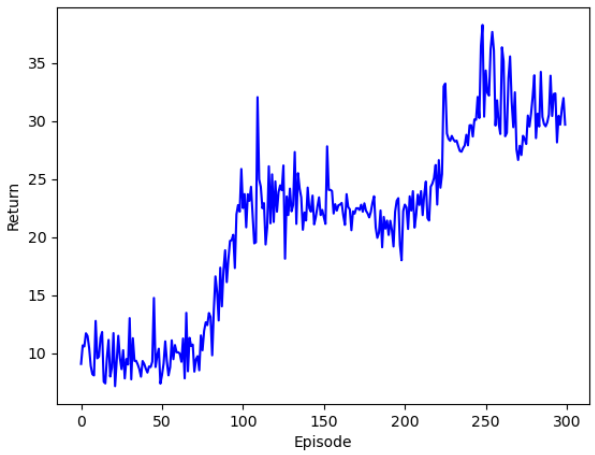

Fig. 5. Recompensas para cada episódio do treinamento.

- Avaliação

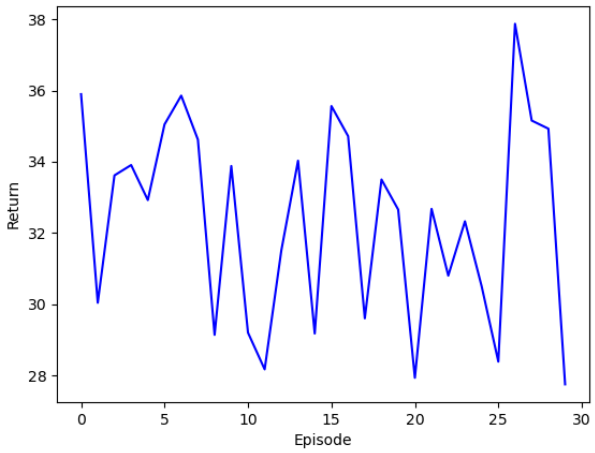

Fig. 6. Valores de avaliação para cada episódio.

- Retorno O retorno médio da avaliação foi de $32,38$, e, ao analizar os tempos de cada episódio, percebeu-se que todos obtiveram sucesso.

Para esta rede, com a primeira recompensa foi possível observar que os valores qe recompensa que mostravam sucesso variavam a partir do seguinte episódio: episode: 240/300, time: 200, score: 27.9175, visto isso, valores acima disso já significavam que o agente cumpriu sua tarefa. Ademais, precebeu-se que durante o treinamento o agente demorou para realizar seu aprendizado, normalizando a recompensa alta a partir dos 200 episódios. Dessa forma, o agente teve um aprendizado lento mas eficaz.

2) Segunda Rede Neural:

- Treinamento

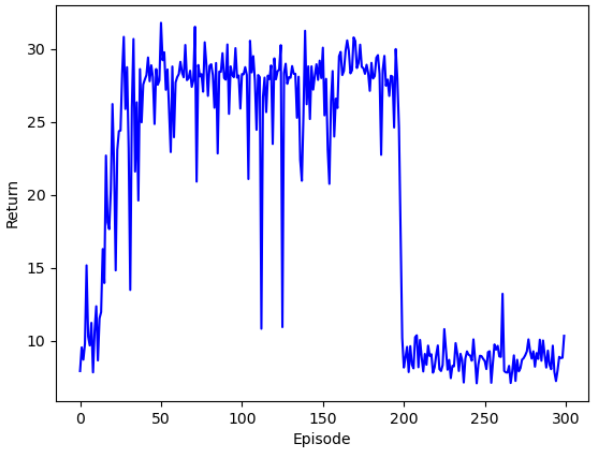

Fig. 7. Valores de recompensa para cada episódio de treinamento.

- Avaliação

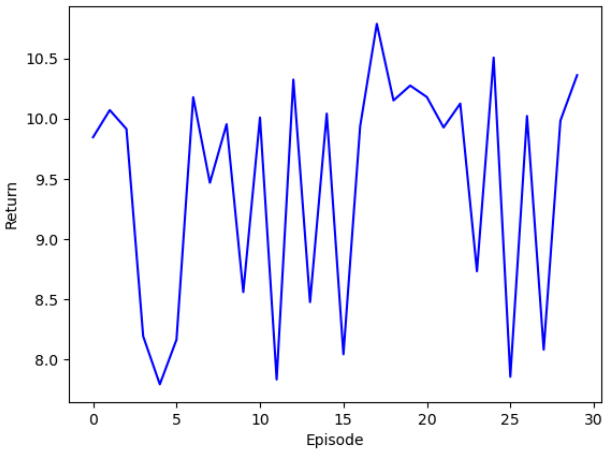

Fig. 8. Valores de recompensa para cada episódio de avaliação.

- Retorno
  O retorno dessa rede neural foi insatisfatório após um treinamento, não conseguindo manter o poste equilibrado em nenhum dos trinta episódios. Obteve-se um valor de 9.4 para o retorno médio.

É visível a diferença gritante entre a avaliação desta duas redes neurais para esta recompensa. A primeira rede neural, mais simples, teve sucesso em treinar o agente no período dado, enquanto a segunda não conseguiu. Quanto ao comportamento do robô houve uma melhora em seu desempenho, visto que permanecia mais no centro do ambiente e com velocidades menores.

C. Recompensa 2

Esse esquema de recompensa foi obtido da Literatura e constava como formulação para a recompensa muito eficaz.

1) Primeira Rede Neural:

- Treinamento

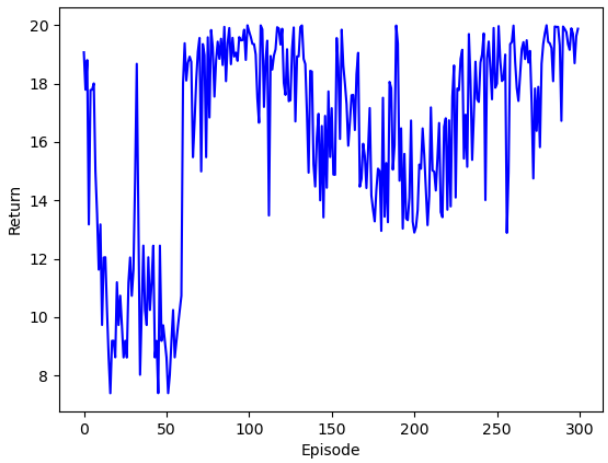

Fig. 9. Gráfico do retorno ao longo dos episódios do treinamento.

- Avaliação

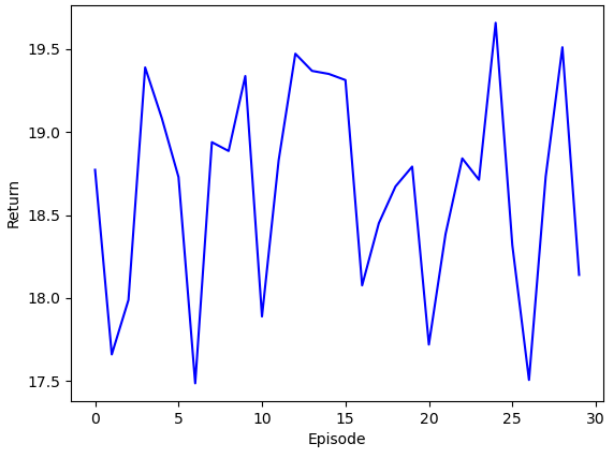

Fig. 10. Avaliação do retorno ao longo dos 30 episódios da avaliação.

- Retorno
  No critério de tempo, todos os episódios apresentaram o tempo máximo de 200 unidades. Quanto ao retorno do *score*, o código retornou média de 18.67

No gráfico do treinamento, nota-se uma clara evolução no aprendizado, uma vez que por volta de 70 episódios o valor médio do retorno aumenta substancialmente, por mais que continue apresentando uma alta variação. Quanto ao gráfico da avaliação, sugere que o algoritmo ainda não convergiu por completo, dado as variações ainda expressivas. Ao analisar o retorno, no entanto, nota-se um resultado ótimo no valor observado do tempo e uma média de *score* muito próxima da convergência.

2) Segunda Rede Neural:

- Treinamento

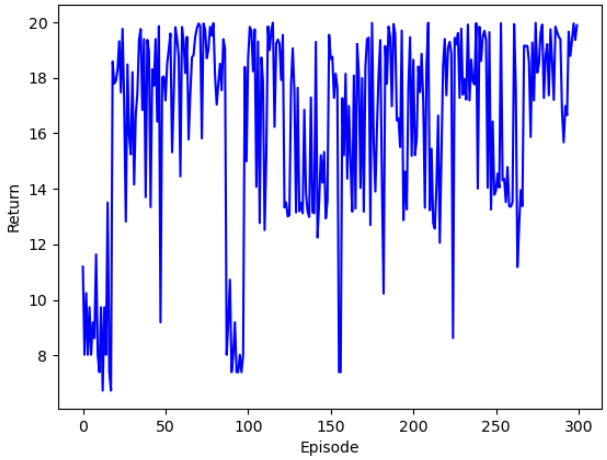

Fig. 11. Gráfico do retorno ao longo dos episódios do treinamento.

- Avaliação

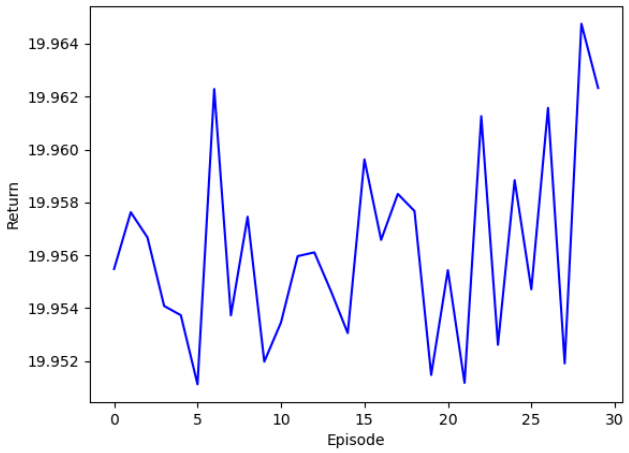

Fig. 12. Avaliação do retorno ao longo dos 30 episódios da avaliação.

- Retorno
  No que se refere ao tempo, o retorno obtido foi de 200 para todos os 30 episódios estudados na avaliação. Quanto ao retorno do *score*, o valor foi de 19.96

No gráfico do treinamento, nota-se uma evolução muito mais rápida dado que antes do episódio 50 já ocorre um aumento drástico no valor dor retorno, indicando que esse modelo de recompensa se adaptou melhor com uma rede neural mais robusta. O gráfico da avaliação confirma essa tendência e, embora o valor do retorno ainda não esteja dentro

dos 70% aceitáveis em geral para a convergência (retorno médio de 21), nota-se que o valor está consideravelmente estável na faixa de 19.956, indicando uma constância que não fora observada facilmente com os outros padrões de recompensa .O retorno também confirma os dados, revelando, além da média esperada do retorno, um valor ótimo para o tempo.

D. Recompensa 3

Baseado no esquema de recompensa 2, o grupo elaborou mais um esquema de recompensa seguindo o que fora abordado no texto estudado da Literatura. Abaixo constam os resultados desse novo padrão de recompensa proposto.

1) Primeira Rede Neural:

- Treinamento

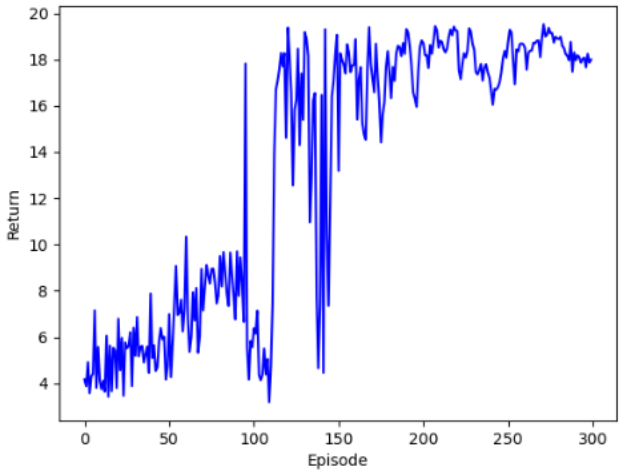

Fig. 13. Gráfico do retorno ao longo dos episódios do treinamento.

- Avaliação

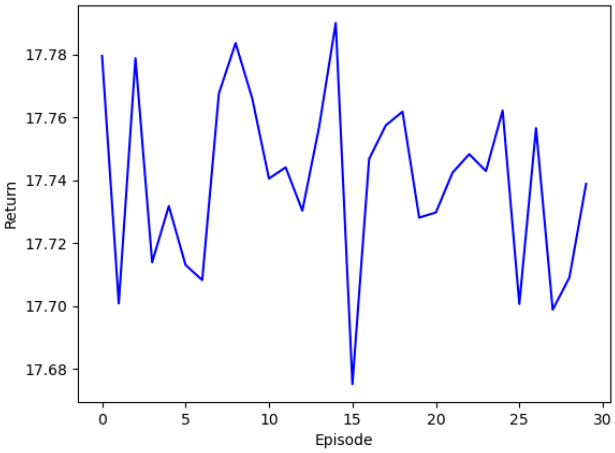

Fig. 14. Avaliação do retorno ao longo dos 30 episódios da avaliação.

- Retorno
  Em termos de tempo, todas os 30 episódios retornaram o valor máximo, 200 unidades. No critério de *score*, o valor médio retornado foi de 17.74

No gráfico do treinamento, nota-se uma clara evolução no aprendizado, uma vez que por volta de 120 episódios o valor médio do retorno aumenta substancialmente. Nota-se, ao realizar a comparação com a mesma rede neural no padrão de recompensa anterior, que esse demorou mais para obter um drástico aumento do retorno. No entanto, uma vez que esse aumento foi alcançado, percebe-se uma menor variação de retorno, Quanto ao gráfico da avaliação, nota-se um padrão que varia pouco, com uma menor variação em relação ao esquema de recompensa anterior com uma mesma rede neural, embora o valor do retorno ainda esteja abaixo do considerado para a convergência. Ao analisar o retorno, percebe-se um valor ótimo para o tempo mas, como esperado, um baixo valor para o retorno.

2) Segunda Rede Neural:

- Treinamento

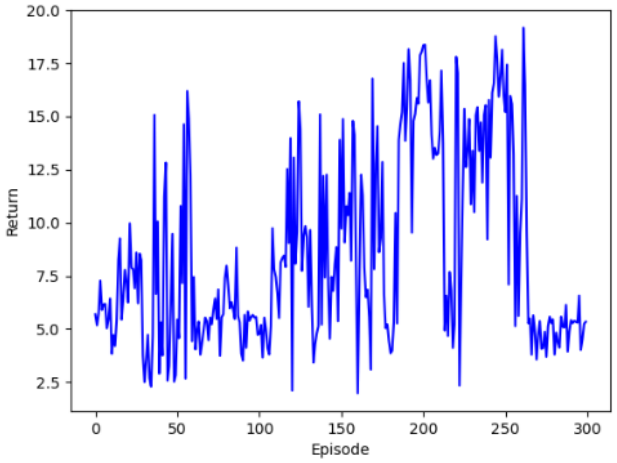

Fig. 15. Gráfico do retorno ao longo dos episódios do treinamento.

- Avaliação

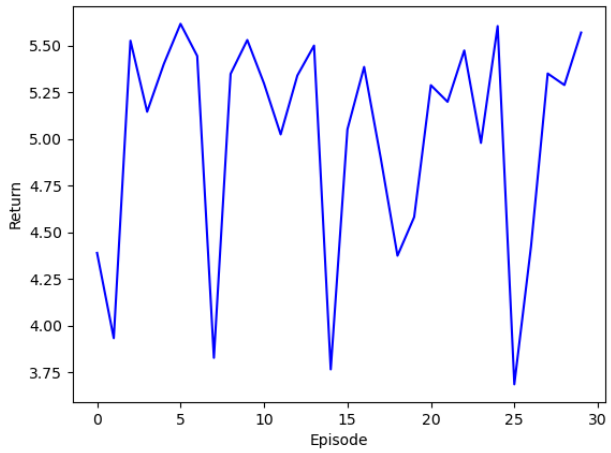

Fig. 16. Avaliação do retorno ao longo dos 30 episódios da avaliação.

- Retorno
  Em termos de tempo, a média dos episódios resultou em 9.03 unidades. No critério de *score*, o valor médio retornado foi de 5.01

No gráfico do treinamento, nota-se que, diferente do que fora observado previamente, o algoritmo não apresenta de forma evidente um aumento drástico do valor do retorno que se mantém razoavelmente constante ao longo do tempo, ficando com a média de retorno previsto bem menor do que nos outros casos analisados. Quanto ao gráfico da avaliação, nota-se uma variação considerável mas com um valor médio muito baixo do retorno. Ao analisar o retorno, percebe-se a confirmação do observado, com um valor de tempo de cerca de 5% do que foi atingido no caso acima e também com todos os casos da recompensa 2. O valor médio também está muito abaixo do que foi previamente obtido em outras combinações. Desse modo, nota-se que essa recompensa apresenta uma dificuldade muito maior de convergência quando submetida a uma rede neural mais complexa.

VI. Conclusões

Dessa maneira, conclui-se que a primeira rede neural é mais eficiente quanto ao treinamento do agente para a resolução do problema CartPole-v0, visto que em todas rewards utilizadas teve desempenho superior. Assim, percebe-se que a profundidade da rede não é correlacionada com sua capacidade de aprendizado, visto que redes mais rasas podem apresentar desempenho melhor para a realização de certas tarefas

Ademais, quanto as recompensas modificadas, é possível ver que as diferentes da padrão conseguiram imprimir um comportamento melhor no agente, fazendo ele se mover menos bruscamente e ficar no centro do ambiente gerado. Além disso, com a segunda recompensa o agente teve o aprendizado mais rápido, visto que antes dos 50 episódios já havi um período estável conseguindo executar a tarefa.

Portanto, este trabalhou conseguiu resolver o problema CartPole, a partir de diferentes reward e redes neurais, além de analisar os efeitos das mudanças de tais elementos.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation (**?**).

References

[1] Solving Open AI's CartPole Using Reinforcement Learning Part-2 Disponível em: https://medium.com/analytics-vidhya/solving-open-ais-cartpole-using-reinforcement-learning-part-2-73848cbda4f1

[2] Infinite Steps CartPole Problem With Variable Reward Disponível em: https://towardsdatascience.com/infinite-steps-cartpole-problem-with-variable-reward-7ad9a0dcf6d0

[3] Deep Learning Book Capítulo 70 Disponível em: https://www.deeplearningbook.com.br/deep-q-network-e-processos-de-decisao-de-markov/

[4] Material de aula do curso CT-213