

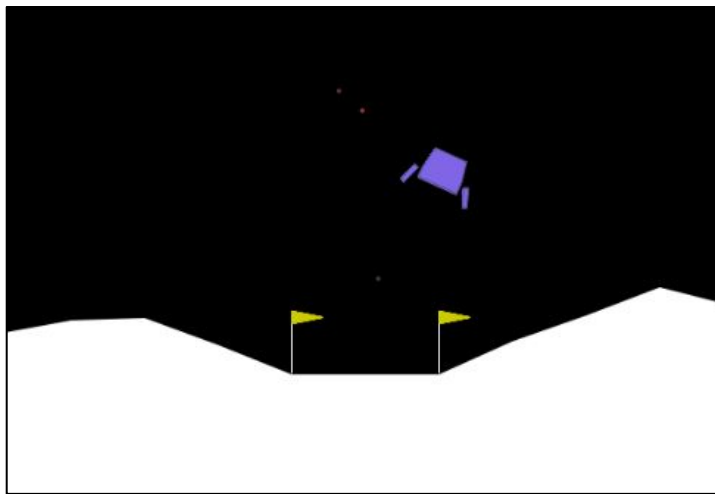


Aprendizado por Reforço

Alunos: Emanuel Mendes e Tiago Cassol

Quais algoritmos foram usados e qual o ambiente de teste

Foi utilizado os algoritmos A2C e PPO para ver qual aprenderia mais rápido e de forma mais eficiente a jogar o Lunar Lander.



Objetivo do lunar lander



O objetivo principal do ambiente é resolver um problema de trajetória de pouso, onde a nave tem que pousar entre as duas bandeiras amarelas sem muito impacto.

Ele possui 4 tipos de ações: 0 - não faz nada, 1 - liga o motor esquerdo, 2 - liga o motor principal, 3 - liga o motor direito.

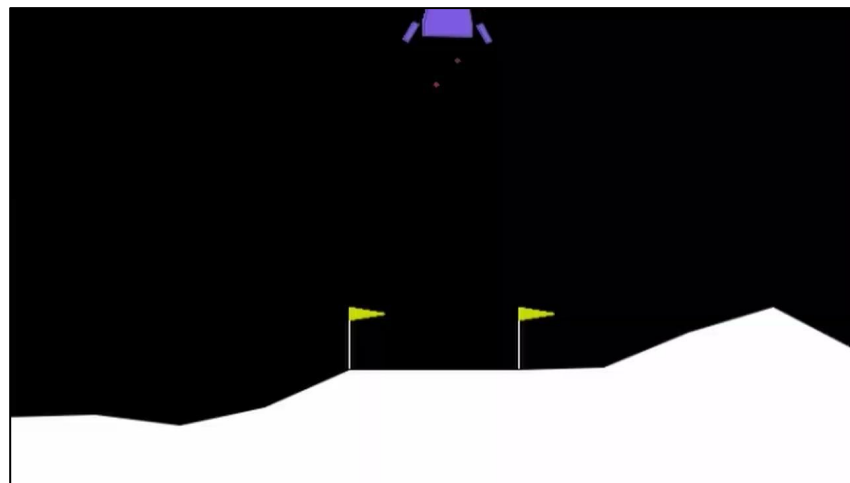
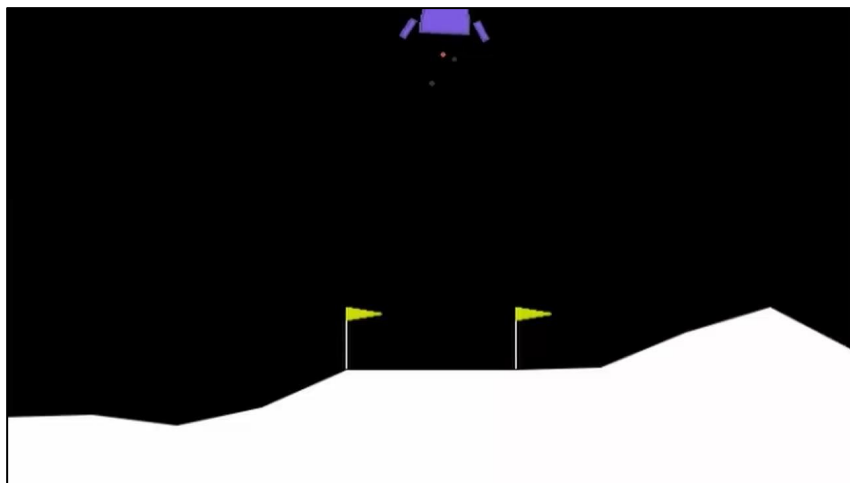
Ele recebe ou perde pontos de acordo com os seguintes requisitos

1. Recebe/perde pontos de acordo com o quão perto/longe a nave está do local de pouso;
2. Recebe/perde pontos de acordo com o quão lento/rápido está se mexendo;
3. Perde pontos quanto mais a nave estiver inclinada;
4. Recebe 10 pontos para cada perna tocando o solo;
5. Perde 0,03 pontos para cada quadro disparado pelos motores laterais;
6. Perde 0,3 pontos para cada quadro disparado pelo motor principal;
7. Recebe/perde 100 pontos por pousar corretamente/bater a nave;

O teste é bem sucedido quando a soma dos pontos é maior que 200.

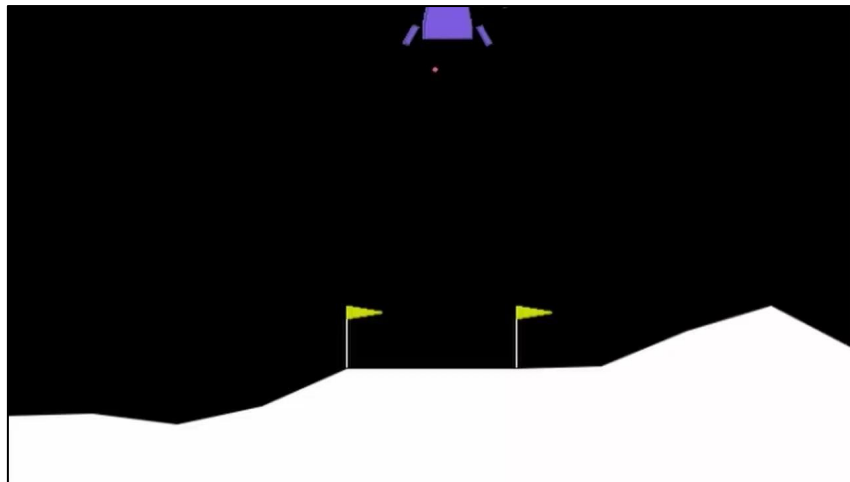
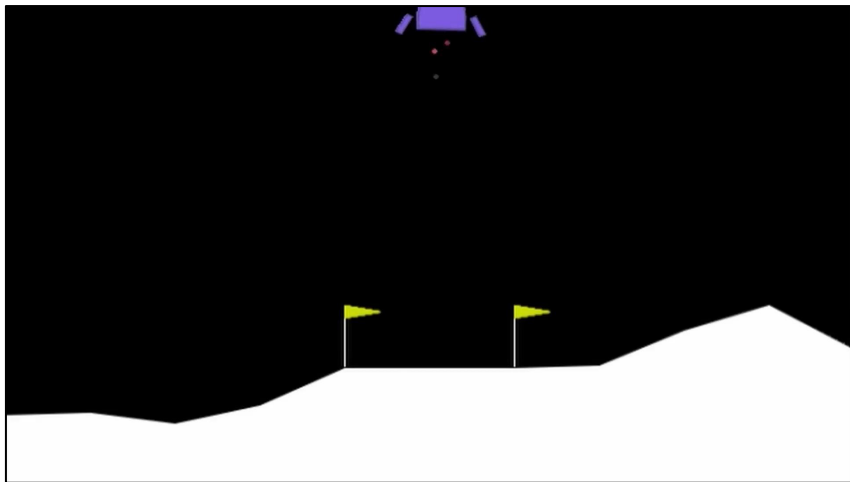
A2C e parâmetros utilizados

Na primeira execução utilizamos 8 environments com uma taxa de aprendizado de 0,00081 e o resultado não foi dos melhores levou 2 minutos e 14 segundos para treinar porém falhava no pouso, foi necessário mais treino e um total de 7 minutos e 40 segundos para que sua execução fosse correta.



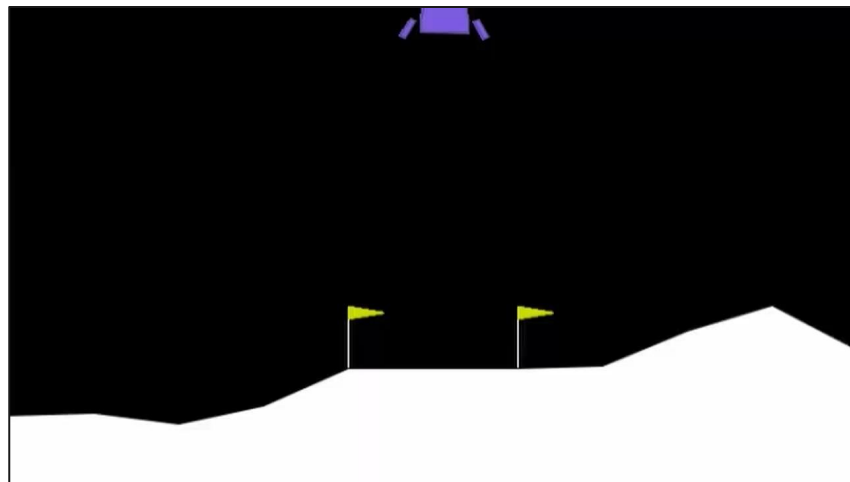
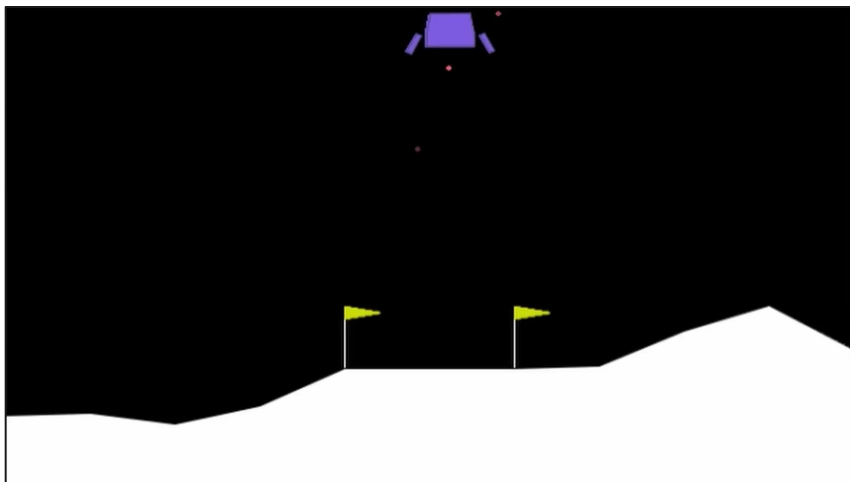
É a vez do PPO treinar

Utilizamos os mesmos parâmetros para treinar o PPO 8 environments e taxa de aprendizado de 0,00081, surpreendentemente o PPO teve uma performance pior, com 2 minutos e 20 segundos para sua primeira execução a nave nem sequer tocava o chão, e após muito treinamento e um total de 6 minutos e 16 segundos ele superou o A2C.



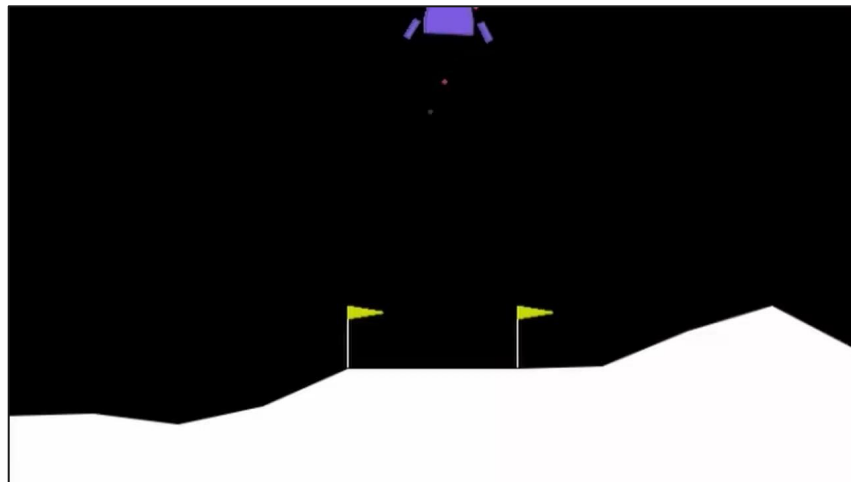
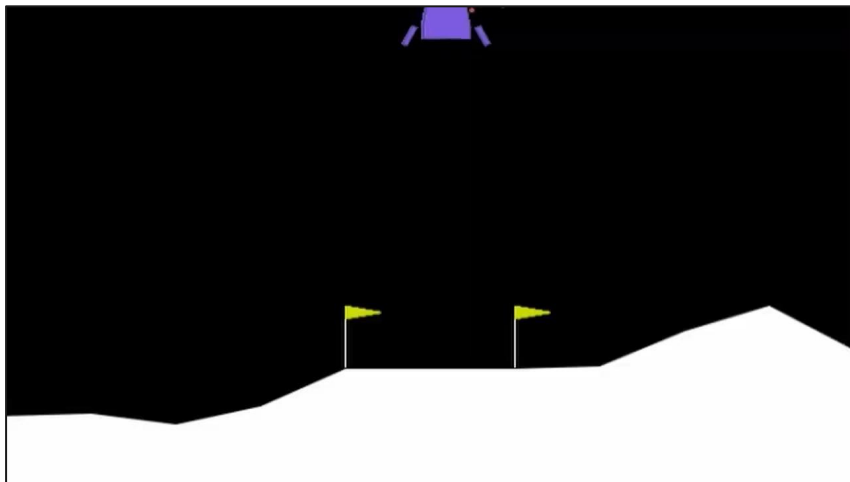
A2C ataca novamente

Testamos novos parâmetros 12 environments e 0,00083 de taxa de aprendizado, foi necessário 2 minutos e 22 segundos para terminar o treinamento, esse era bipolar, ou ele batia com muita força no chão ou flutuava até o tempo acabar. Com 3 minutos e 20 segundos de treinamento ele aprendia a pousar corretamente.



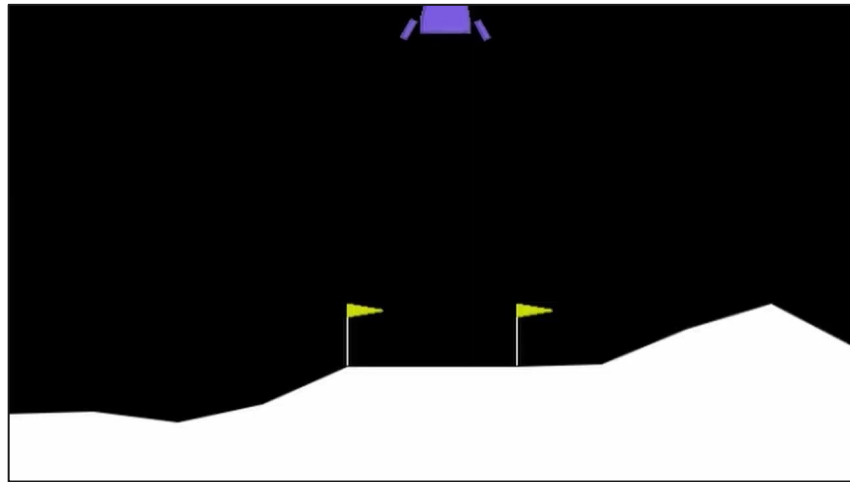
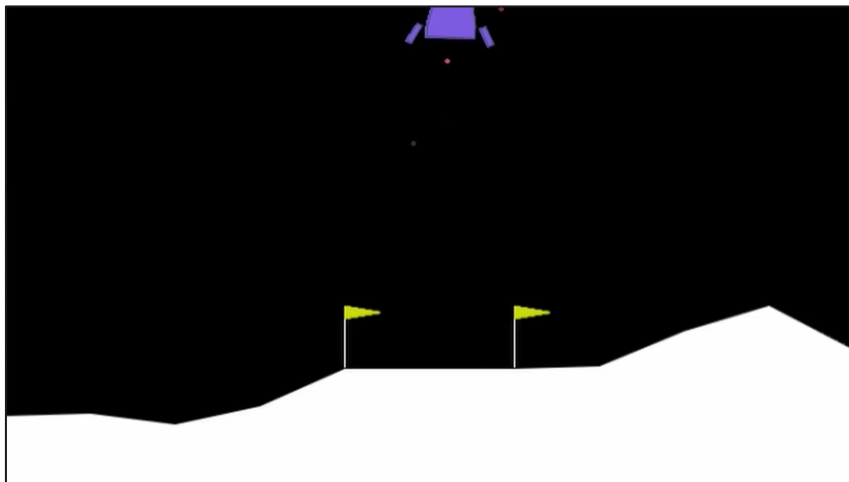
O contra ataque de PPO

Os mesmos parâmetros foram utilizados no PPO e ele precisou de 2 minutos e 21 segundos para treinar, enquanto o A2C treinava pouso, o PPO decidiu que treinaria a fuga e saiu voando para longe da área de pouso. Ele precisou treinar por 5 minutos e 47 segundos pra conseguir perder o medo de pouso.



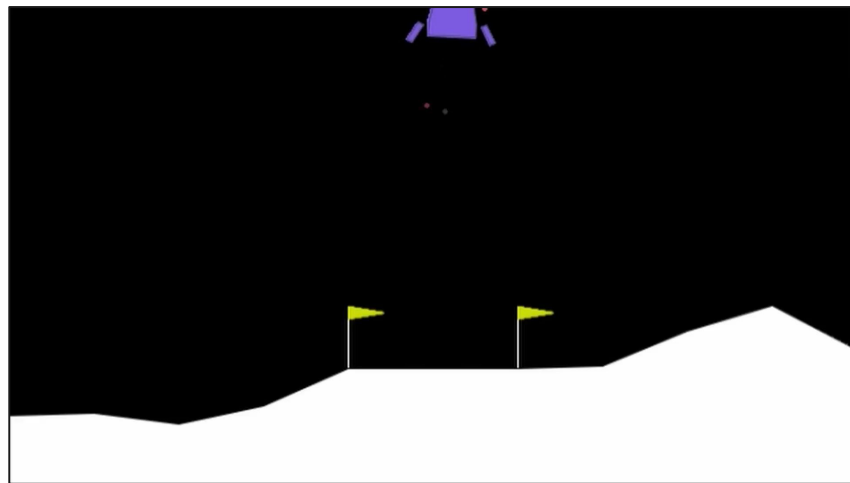
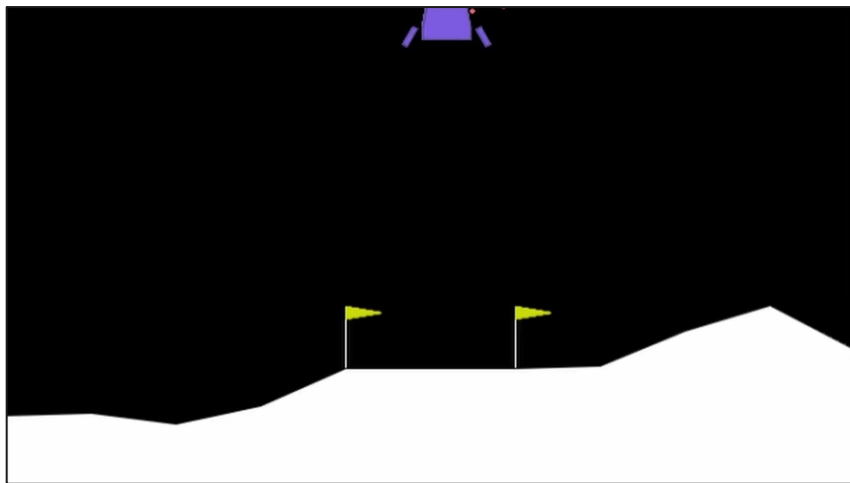
Último round A2C vs PPO

Como última combinação de parâmetros foram testados 16 environments e sua taxa de aprendizado era de 0,00078, ele treinou em bem pouco tempo, apenas 2 minutos e 19 segundos, porém sua execução poderia ser melhor, ele precisou de 4 minutos e 7 segundos para aprender a pousar de forma correta.



(sem título, to sem criatividade pra isso)

Eis o último teste do PPO, com os mesmos parâmetros ele precisou de 2 minutos e 11 segundos para concluir o treinamento, mas acho que ele treinou ouvindo Tokyo Drift, pousou e saiu deslizando, porém com um pouco mais de treino ele conseguiu pousar corretamente.



Considerações finais

Será que Naruto estava certo quando disse que usar clones é melhor para aprender?

A resposta é sim, usar mais environments, com uma taxa de aprendizado mais baixa acaba por ser melhor do que o inverso.

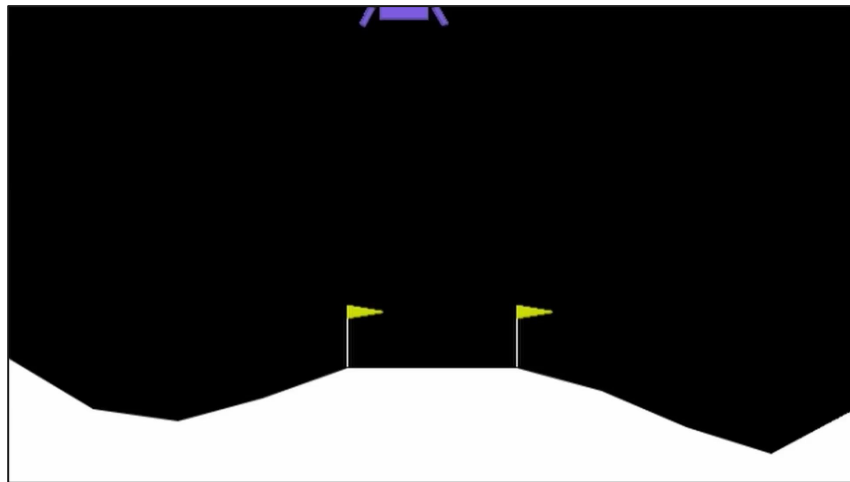


A2C inimigo da gravidade



Lembrem-se de algo, nunca pessa pro A2C fazer um pouso forçado, ou talvez você não sobreviva a queda.

Fonte: baseado em fatos reais.



Mas afinal quem venceu a batalha



Num geral, o A2C se saiu melhor, ele aprendeu em menos tempo que o PPO com 12 e 16 environments, com 8 environments o PPO se saiu melhor. Para executar o lunar lander o melhor algoritmo foi o A2C. Porém o PPO garantiu mais risadas com os erros bobos que ele cometia.


Resultados de testes



A2C (8 environments, taxa de aprendizado de 0,00081 e 200000 timesteps): Primeiro treino levou 134 segundos para executar porém pousava errado, então foi treinado novamente 216 segundos de treino e ele ainda pousava errado ou as vezes nem pousava, com 297 segundos de treino o erro ainda persistia e ele ainda pousava errado ou não pousava, 383 segundos ele aprendeu a pousar mas às vezes ainda ficava flutuando, e com 460 segundos ele aprendeu a pousar corretamente.

A2C (12 environments, taxa de aprendizado de 0,00083 e 200000 timesteps): Foram necessários 142 segundos para um primeiro treinamento ele batia ou demorava demais para pousar, com 200 segundos ele aprendeu de forma correta.

A2C (16 environments, taxa de aprendizado de 0,00078 e 200000 timesteps): 139 segundos e ele errava os pousos ou não descia, 193 segundos de treino ele batia com muita força no chão, foi preciso 247 segundos de treino para ele aprender a pousar.



PPO(8 environments, taxa de aprendizado de 0,00081 e 200000 timesteps): Foi necessário 140 segundos para o primeiro treinamento porém ele voava para fora da tela, 263 segundos de treinamento ele começou a aprender a pousar porém pousava errado, com 376 segundos de treinamento ele pousou corretamente.

PPO (12 environments, taxa de aprendizado de 0,00083 e 200000 timesteps): O primeiro treinamento levou 141 segundos porém ele fugia para longe do local de pouso, com 245 segundos ele flutuava em cima do local de pouso porém não descia, precisou de 347 segundos pra aprender a pousar.

PPO (16 environments, taxa de aprendizado de 0,00078 e 200000 timesteps): 131 segundos para o primeiro treinamento porém ele deslizou para fora da área de pouso, com 237 segundos ele ficava apenas flutuando em cima da área de pouso, com 357 segundos ele ficava em cima da área de pouso porém não descia, e com 465 segundos ele aprendeu a pousar.