Model-Based Reinforcement Learning for Atari

Professor: Tiago Maritan Ugulino de Araújo
Aluno: Lucas da Silva Cruz

Aprendizado por Reforço Baseado em Modelos para Atari

Autores

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, Henryk Michalewski

Para acesso ao artigo https://arxiv.org/abs/1903.00374

Motivação

- No aprendizado por reforço algoritmo com base em modelo vem alcançando resultados espetaculares, mas requer muita interação com o ambiente.
- O projeto baseado em modelo é melhor que pela complexidade da amostra, mas pode existir um modelo para entrada visual?

Motivação

- A ideia é com quão poucos dados você precisa para ensinar um agente razoável para jogar Atari.
- Um modelo por exemplo de 3 agentes são bons mas podem acabar precisando de milhões de quadros para interagir direto.
- E levando em consideração ao ser humano mesmo que não se torne um especialista no jogo, precisaria de 5 minutos para aprender a jogar.

Resumindo

Foi treinado pela equipe uma nova arquitetura para o modelo de previsão de vídeo a ser usado em uma RL baseada em modelo para o Atari.

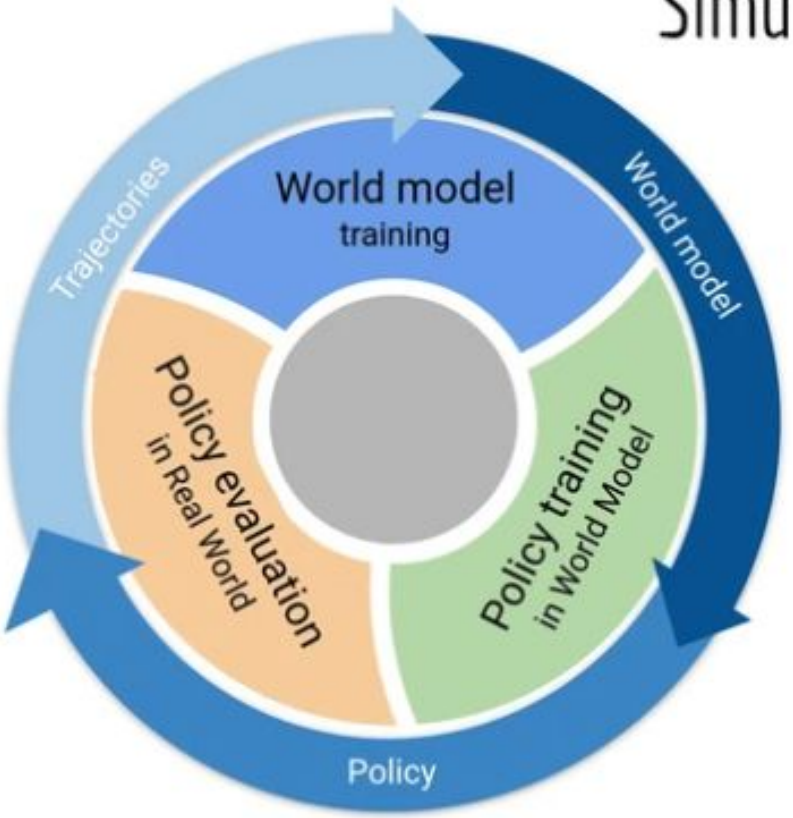
Trazer a possibilidade de "forçar" os agentes a "forçar" o algoritmo.

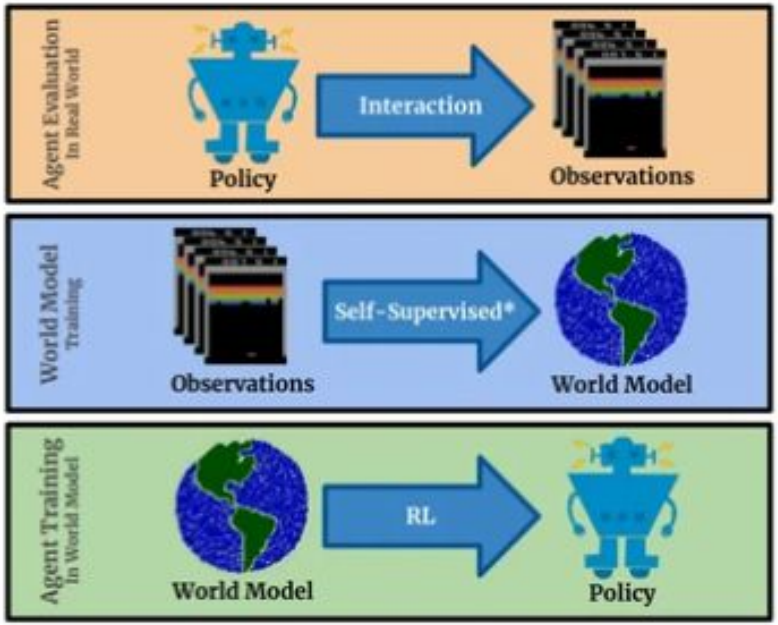
Sendo utilizado poucos dados do ambiente real e sendo observado um desempenho razoável para jogar Atari.

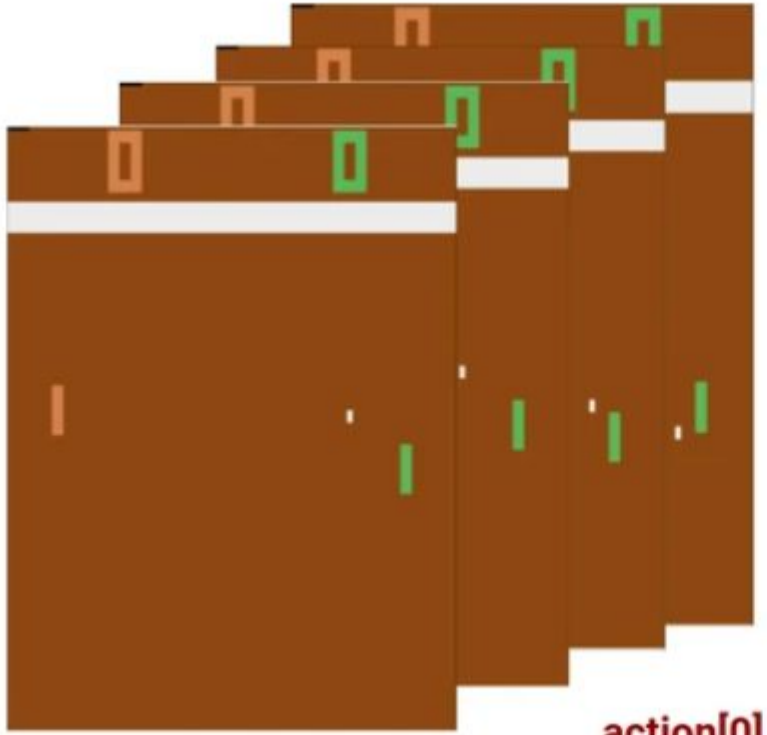
Mas porque atari?

- Pode ser que Atari não venha ser tão importante por ser um simulador leve ou de outra forma onde a computação é barata.
- Mas se pensar em aplicativos, por exemplo, que envolvam veículos autônomos a coleta de dados reais é considerado caro poderia ser uma aplicação.
- Desenvolver métodos que usem menos com uma quantidade ainda menor de dados para treinamento seria algo importante.

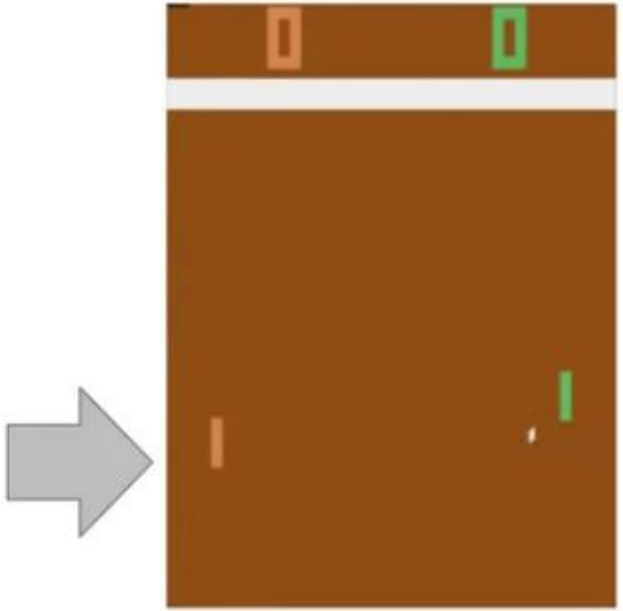
Simulated Policy Learning (SimPLe)

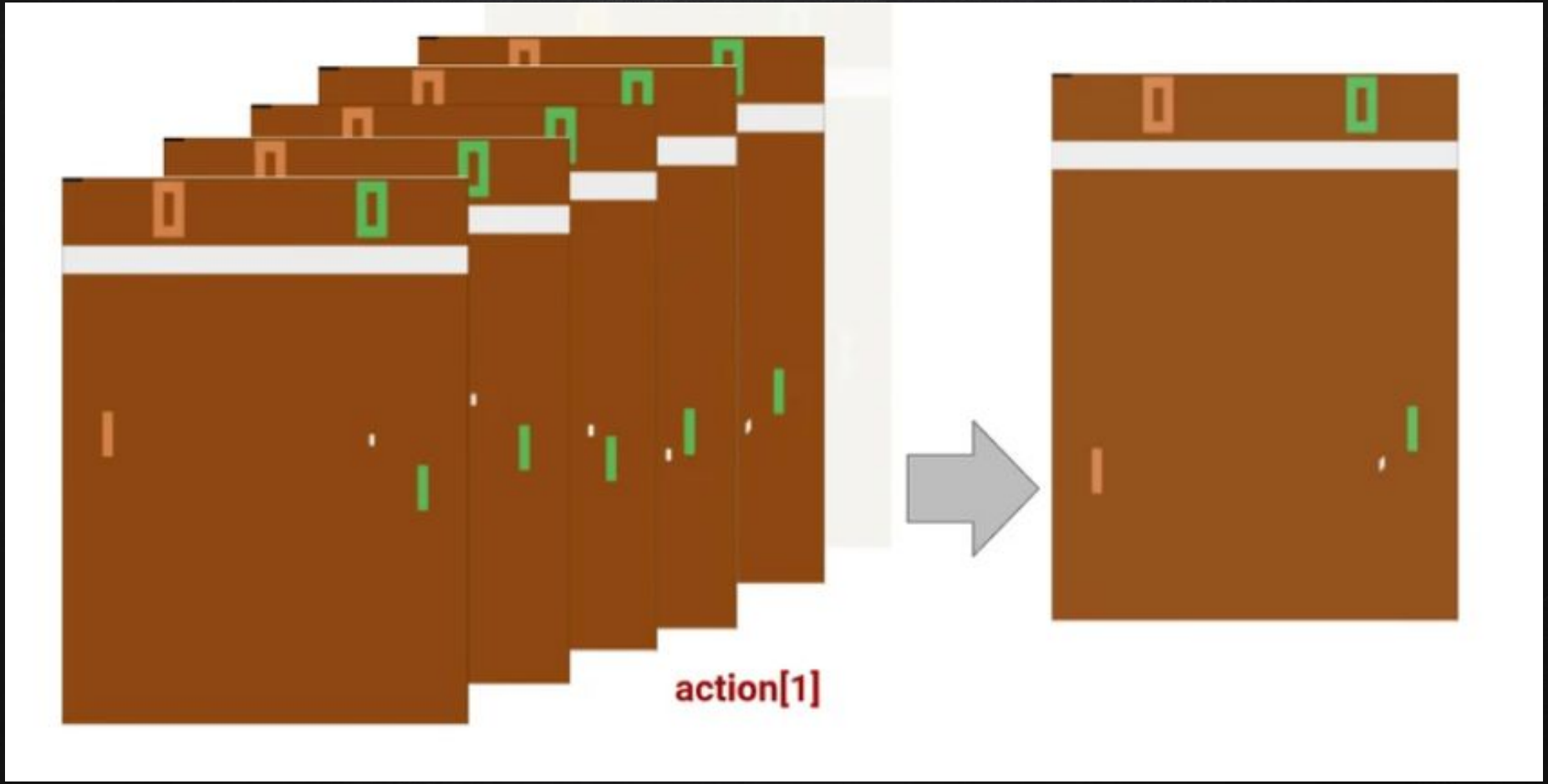


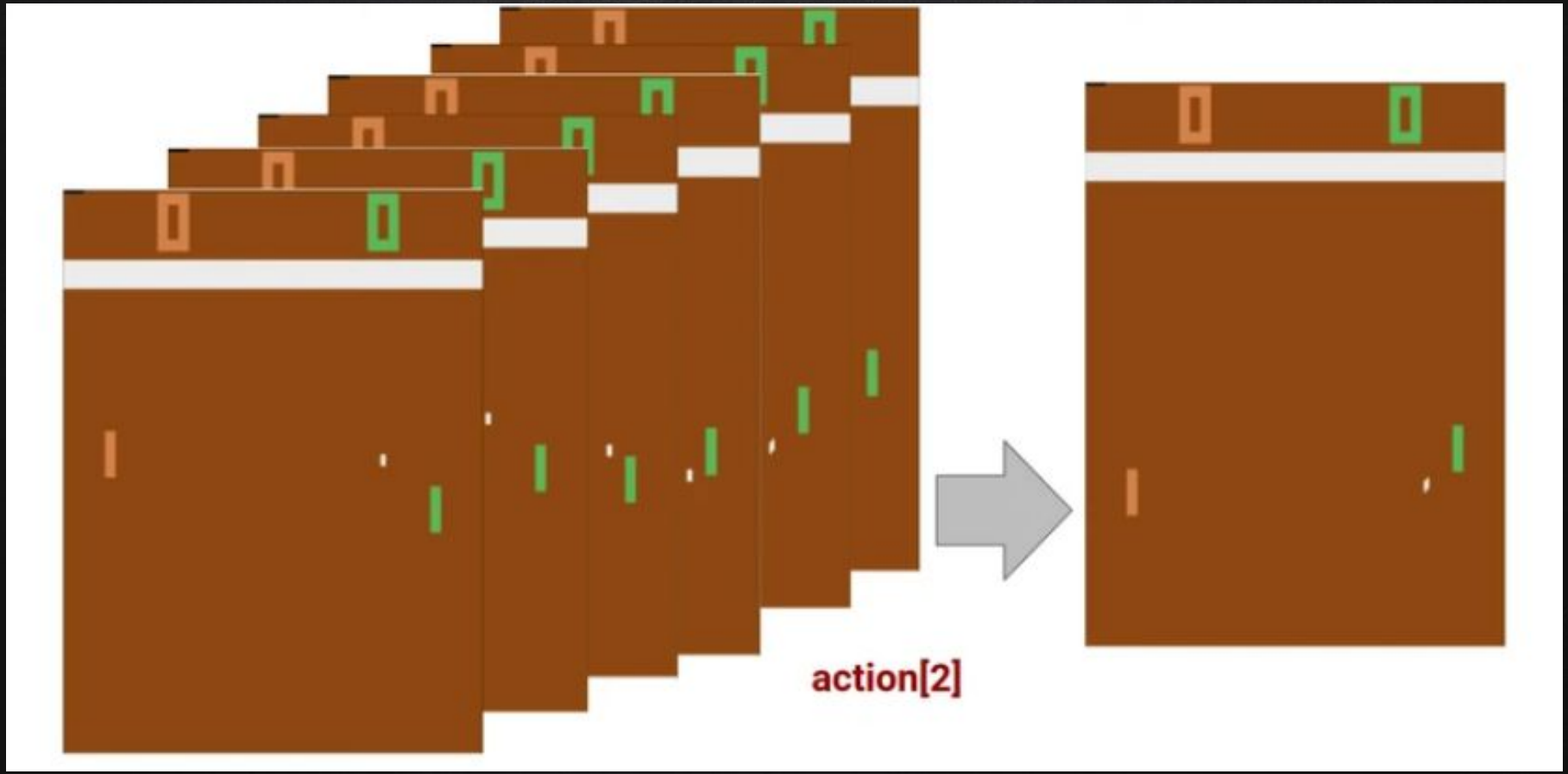




action[0]

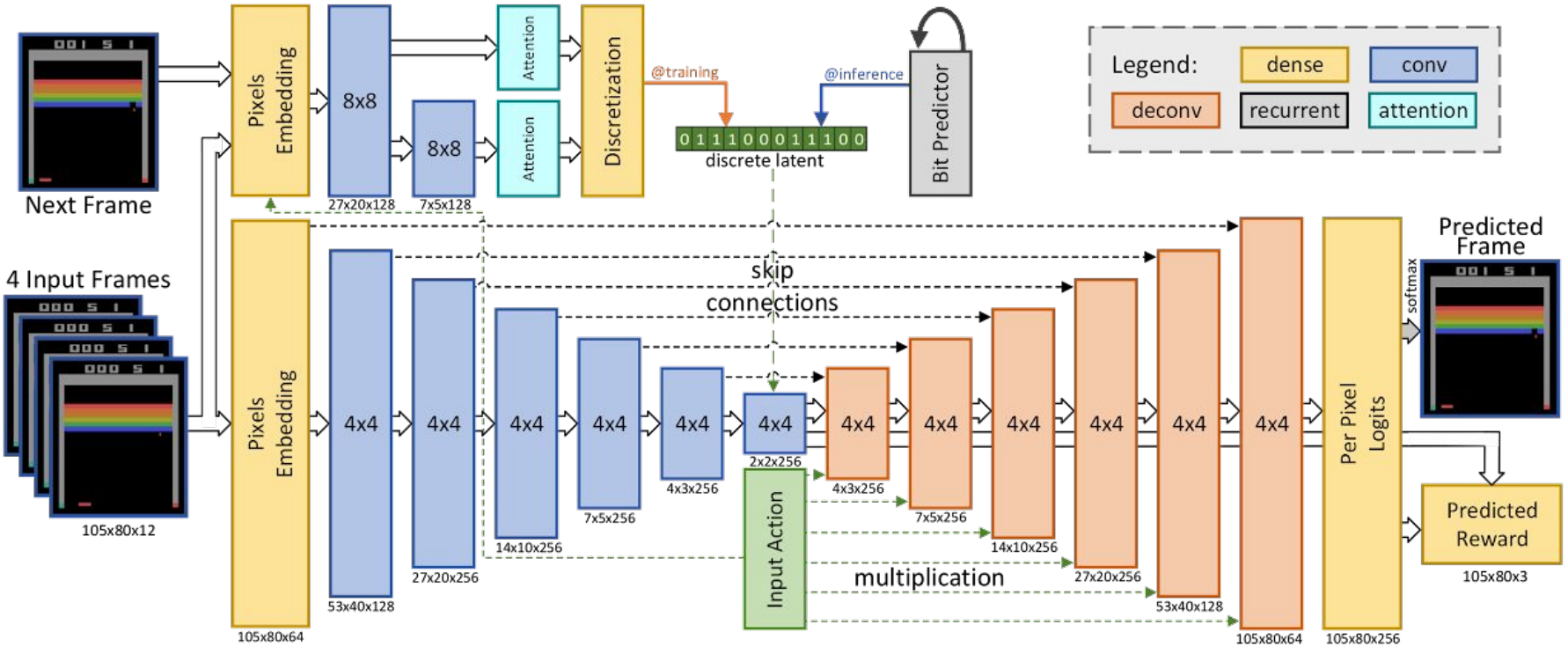



action[1]


action[2]

Modelo discreto estocástico

- O agente aprende com observações brutas de pixel geradas por um modelo de previsão de vídeo;
- A recompensa é prevista com base na representação do gargalo;
- Entre as arquiteturas experimentadas foi verificado que o modelo é uma rede neural convolucional feedforward, que codifica uma sequência de quadros de entrada usando uma pilha de convoluções e dada uma ação realizada pelo agente, decodifica um próximo quadro usando uma pilha de desconvoluções.


Next Frame
Pixels Embedding
8x8
27x20x128
8x8
7x5x128
Attention
Attention
Discretization
@training
@inference
0 1 1 1 0 0 0 1 1 1 0 0
discrete latent
Bit Predictor
Legend:
dense
conv
deconv
recurrent
attention
4 Input Frames
105x80x12
Pixels Embedding
105x80x64
4x4
53x40x128
4x4
27x20x256
4x4
14x10x256
4x4
7x5x256
4x4
4x3x256
4x4
2x2x256
skip
connections
Input Action
multiplication
4x4
4x3x256
4x4
7x5x256
4x4
14x10x256
4x4
27x20x256
4x4
53x40x128
4x4
105x80x64
Per Pixel Logits
105x80x256
softmax
Predicted Frame
Predicted Reward
105x80x3

- Foi descoberto que a introdução da estocasticidade no modelo tem um efeito benéfico.
- Permitindo que a política experimente um conjunto mais diversificado de cenários durante o treinamento.
- Fazendo isso foi adicionado uma variável latente, cujas amostras são adicionadas à representação do gargalo.
- E variáveis discretas funcionam melhor na configuração, codificadas como sequências de bits.

Toda a arquitetura é uma reminiscência de um autoencoder variacional, onde o posterior sobre a variável latente é aproximado com base em toda a sequência…

(quadros de entrada + quadro de destino)...

Onde um valor é apresentado desse posterior usado, juntamente com os quadros de entrada e a ação, para prever o próximo quadro.

Durante a inferência, os códigos latentes são gerados por uma rede LSTM autorregressiva.

Política de treinamento

- É usado o modelo do mundo como simulador imperfeito do ambiente real;
- Ocorre o treinamento dentro do modelo do mundo não sendo afetado a complexidade da amostra
- Ocorre o lançamentos do modelo mundial onde tendem a se degradar após muitas etapas

Inícios aleatórios

- Inicia etapas em pontos aleatórios das trajetórias do ambiente real
- os lançamentos do simulador podem ser relativamente curtos, mas o agente ainda pode aprender o jogo completo

Resultados da pesquisa

O objetivo principal do artigo era usar métodos baseados em modelos para obter eficiência de amostra de última geração.

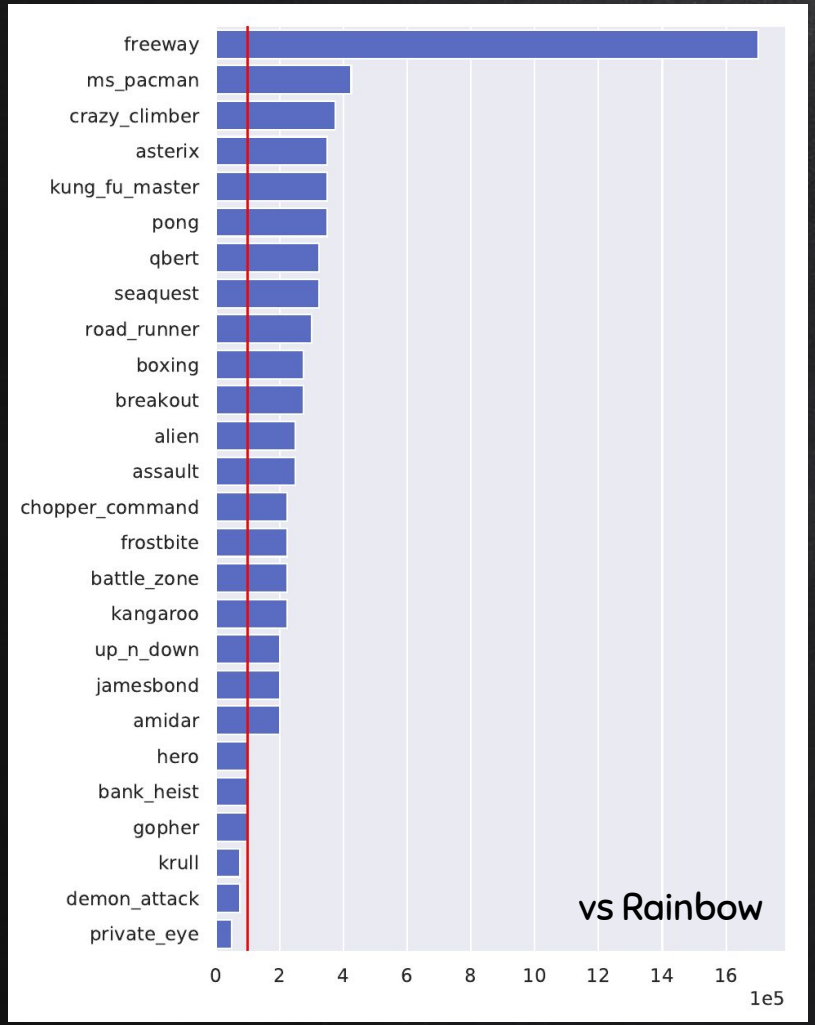
Para isso, a equipe rastreou a resposta à seguinte questão.

"Que pontuação podemos alcançar dentro do orçamento modesto de 100 mil interações (cerca de 2 horas de jogo em tempo real)?"

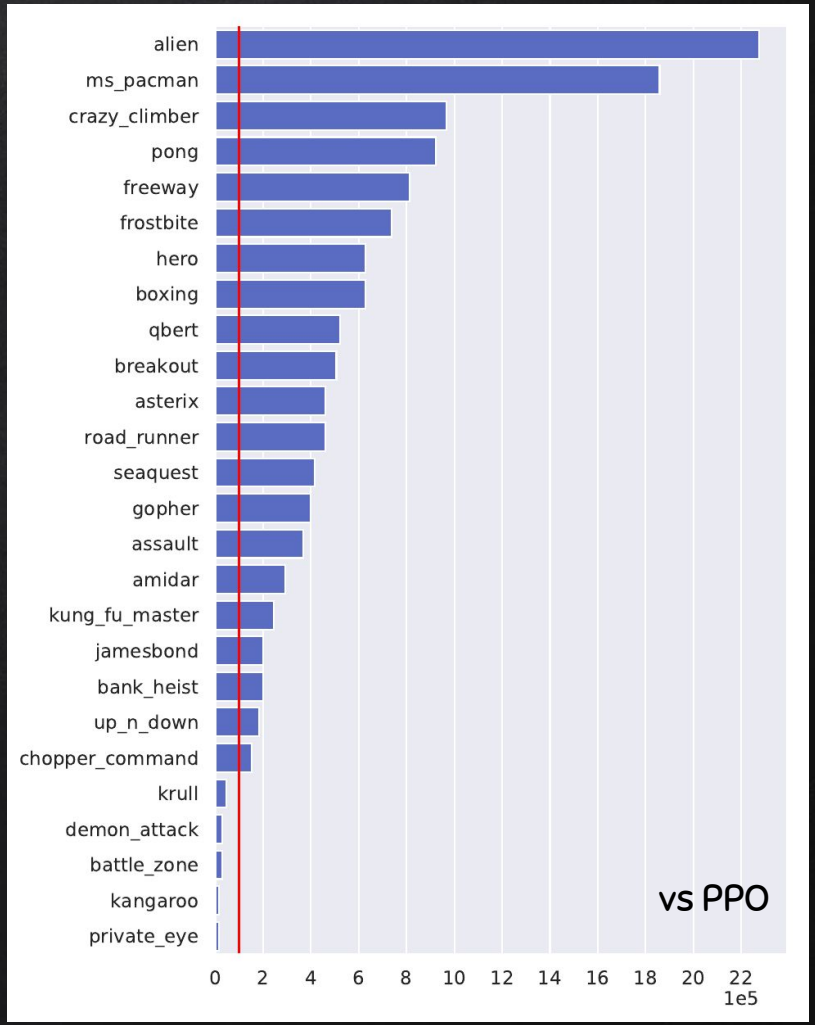
Resultados da pesquisa

Para encontrar a resposta foi comparado com o método Rainbow, algoritmo sem modelo de última geração para jogos da Atari, reajustado para desempenho ideal usando 1 milhão de interações com o ambiente.

E em outro momento foi feito uma comparação com a implementação PPO, que segue a ideia de atualizar a política diretamente para aumentar a probabilidade de ações que proporcionem uma recompensa futura maior



LINHA VERMELHA INDICA O NÚMERO DE INTERAÇÕES



O que foi conquistado

Jogos Resolvidos

Jogos Perfeitos para Pixels

Novo Modelo com alto grau de eficiência

O que foi identificado

Obrigado!

Dúvidas?

PRESS START

Referência bibliografica

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., ... & Michalewski, H. (2019) v4. Última revisão em 19 Fev 2020. Model-based reinforcement learning for atari. arXiv preprint arXiv:1903.00374.