AGL10225 – Aprendizado por Reforço

Planejamento de rota de veículo articulado em ambiente de estacionamento

DAYANA CARDOSO

WALTER FRANK

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Introdução

O planejamento de movimento para veículos articulados (ex: uniciclo com reboque) em ambientes não estruturados enfrenta o desafio da **complexidade cinemática local e topológica global.**

- Controladores baseados em otimização (MPC) resolvam a estabilização local, mas sofrem em navegação global com **obstáculos** (mínimos locais).

- Abordagens de Aprendizado por Reforço (RL) prometem robustez global, mas sofrem com o problema do **horizonte longo**

Nesse contexto, exploramos o potencial do Heuristic Guided Reinforcement Learning (HuRL) para **reduzir a esparsidade** da recompensa e aumentar a eficiência de treinamento de agente de RL.

HuRL

Cheng et al, introduziram o **Heuristic-Guided Reinforcement Learning (HuRL)** como uma solução para acelerar o RL. O HuRL utiliza uma heurística (mesmo que imperfeita) para "encurtar" o horizonte efetivo do problema através da moldagem de recompensa (reward shaping) baseada em potencial. No entanto, a eficácia do HuRL depende criticamente da qualidade desta heurística inicial.

Nesse contexto, exploramos diferentes **heurísticas** para modelar a recompensa do agente, comparando seu impacto efetivo na eficiência e velocidade de treinamento.

Metodologia

Modelagem do Domínio Físico

Modelamos a entidade trator-trailer, seu ambiente físico, interações e dinâmicas em um ambiente da api Gymnasium customizado.

Definição das Heurísticas para Reward Shaping

Instrumentamos o ambiente de aprendizado com duas heurísticas principais, que modelam a recompensa do agente a cada passo de forma distinta:

- **Distância euclidiana:** Recompensa o agente por reduzir a distância em linha reta até o objetivo
- **Distância de menor caminho considerando obstáculos (BFS):** Recompensa o agente por seguir o menor caminho válido até o objetivo, reduzindo a menor distância calculada pelo BFS.

Metodologia

Treinamento e Comparação

- Treinamos agentes de aprendizado por reforço por um número fixo de passos com cada heurística, usando o mesmo algoritmo
- Comparamos o desempenho de treinamento dos agentes entre si e ao baseline - um agente treinado sem heurística.

Modelagem do Domínio

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Pilha de Software

Python

Linguagem de programação principal

OpenAI Gymnasium

API para ambiente de aprendizado por reforço

Stable-baselines3

Implementações de algoritmos de RL compatívies com Gymnasium

Tensorboard

Acompanhamento em tempo real de treinamentos

CasADi

Otimização numérica, diferenciação algoritmica

Outras bibliotecas

Jupyter Notebooks, MatplotLib, Pytest, Numpy

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

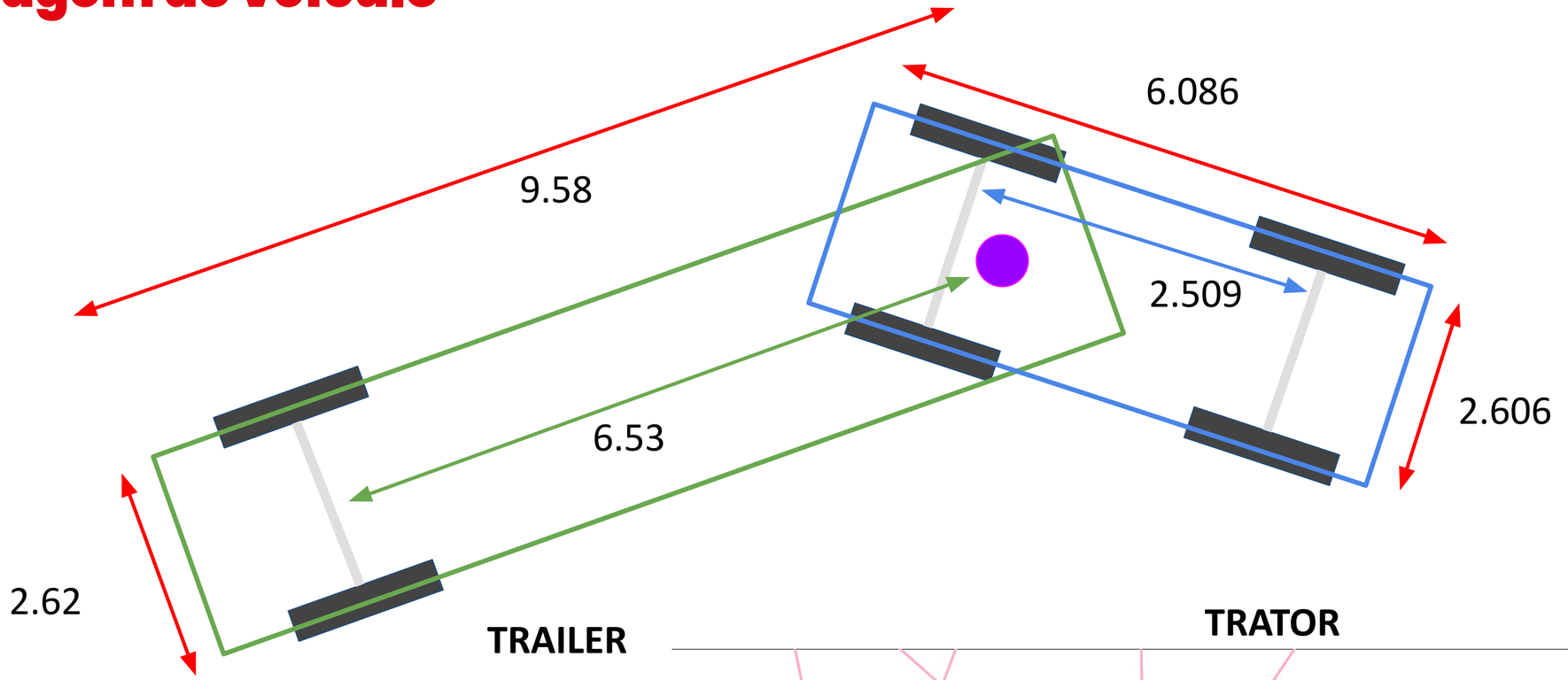
.inf
INSTITUTO DE INFORMÁTICA
UFRGS

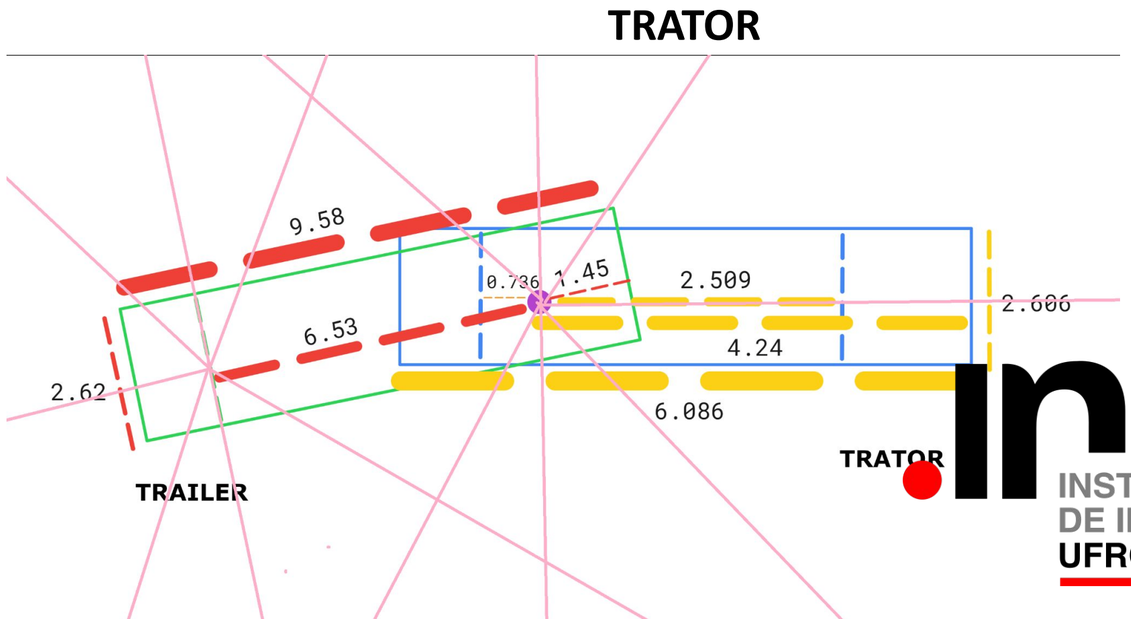
O veículo

Representamos o veículo como um sistema cinemático de bicicleta articulado com três componentes principais:

- Trator: posição (x, y), orientação θ, eixo traseiro como referência
- Trailer: acoplado ao trator pela quinta roda, orientação $\theta\_trailer = \theta - \beta$
- Articulação: ângulo β entre trator e trailer

Modelagem do veículo





UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

INF
INSTITUTO
DE INFORMÁTICA
UFRGS

Equações Cinemáticas - Modelo Bicicleta Cinético

O sistema é não-holonômico, o que significa que o veículo não pode se mover instantaneamente em qualquer direção (ele não pode andar de lado, por exemplo).

- Posição do trator

$$\dot{x} = v\cos(\theta_0) \qquad \dot{y} = v\sin(\theta_0)$$

- Orientação do trator

$$\dot{\theta}_0 = \frac{v}{L_1}\tan(\delta)$$

- Orientação do trailer

$$\dot{\theta}_1 = \frac{v\sin(\theta_0 - \theta_1) - d\cdot\dot{\theta}_0\cos(\theta_0 - \theta_1)}{L_2}$$

Onde:

x, y	Posição global do trator
v	Velocidade logitudinal do trator
δ	Ângulo de esterçamento do trator
L_1	Distância entre-eixos do trator
L_2	Distância do ponto de engate até o eixo do reboque

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS

Ambiente físico

Mapa

- Pares de fileiras de vagas apontando em direções opostas, com paredes entre elas

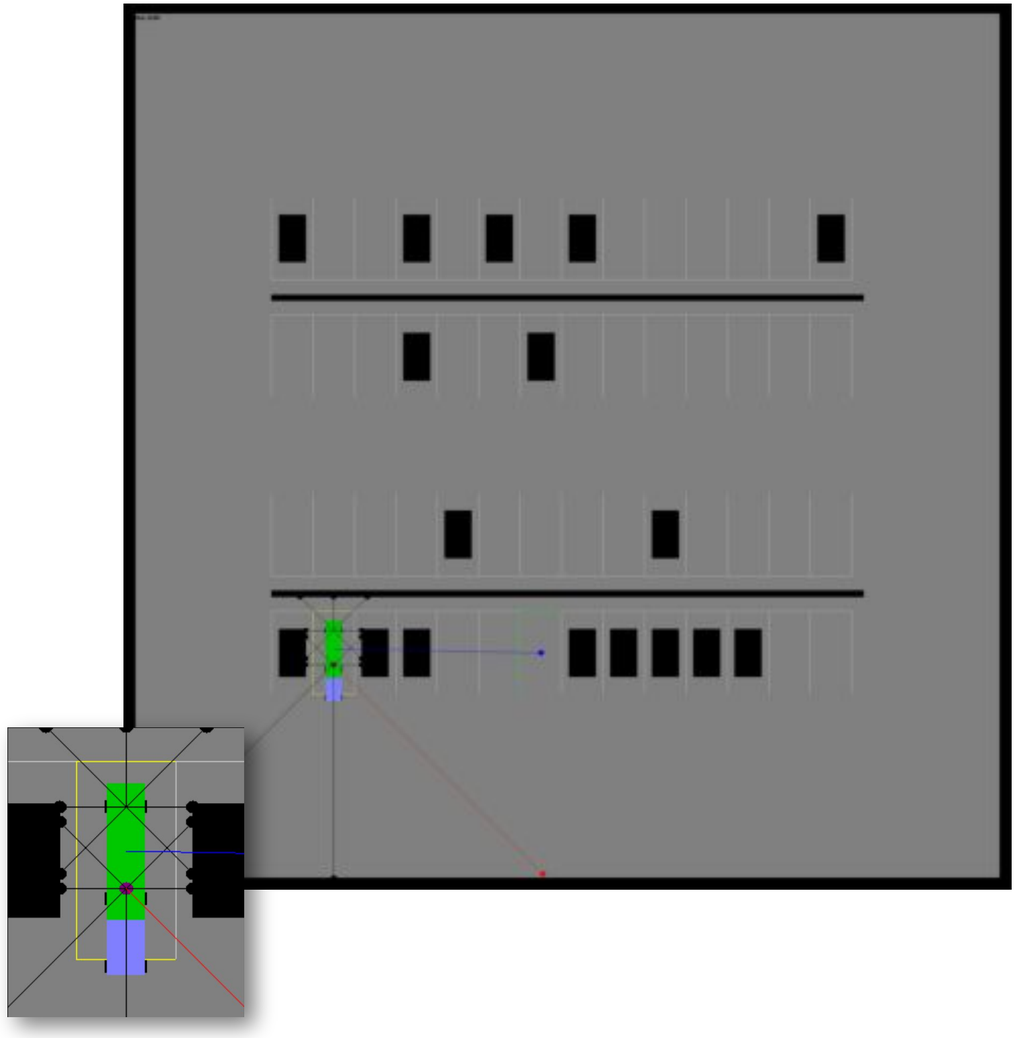
Dimensões

- 150m x 150x - 22500m^2

Randomização de domínio

- A cada geração, uma vaga aleatória é escolhida como ponto de partida e outra é escolhida como ponto de chegada (alvo)
- Quaisquer outras vagas tem uma chance de 25% de possuir um veículo estacionado sobre elas (obstáculo)

Ambiente



UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Ambiente Gymnasium

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Espaço de observação

$$\mathbf{s}_t = [v, \theta, \beta, \alpha, \mathbf{r}, \rho_g, \psi_{rel}, \Delta\theta_{tr}, \Delta\theta_{tl}]^T$$

Onde:

v, α : Velocidade longitudinal e ângulo de esterçamento atuais do trator;

θ, β : Orientação global do trator e ângulo de articulação do reboque;

$\mathbf{r} \in [0, 1]^{14}$: Vetor de leituras de sensores LIDAR (\textit{raycasts}) para detecção de obstáculos;

ρ_g : Índice de proximidade ao alvo, dado por $(1 + ||\mathbf{p} - \mathbf{p}_{goal}||_2)^{-1}$

ψ_{rel} : Ângulo de azimute relativo (egocêntrico) em direção ao alvo;

$\Delta\theta_{tr}, \Delta\theta_{tl}$: Erros de orientação do trator e do reboque, respectivamente, em relação à pose final desejada.

Função de Recompensa (esparsa)

+100 por completar o objetivo (estacionar na vaga de destino)

+100 por alinhar o trailer na mesma orientação da vaga

-100 por colisão com paredes ou outros obstáculos

-100 por canivete (trailer acima do ângulo crítico)

-20 distribuídos ao longo do episódio como penalidade por tempo

-0.1 por velocidade zero a cada passo(penalidade por ficar parado)

-0.02 por mudança brusca de esterçamento (smoothness penalty)

Fator da heurística

+1.0 por progresso em direção ao objetivo (determinado pela heurística)

Algoritmo

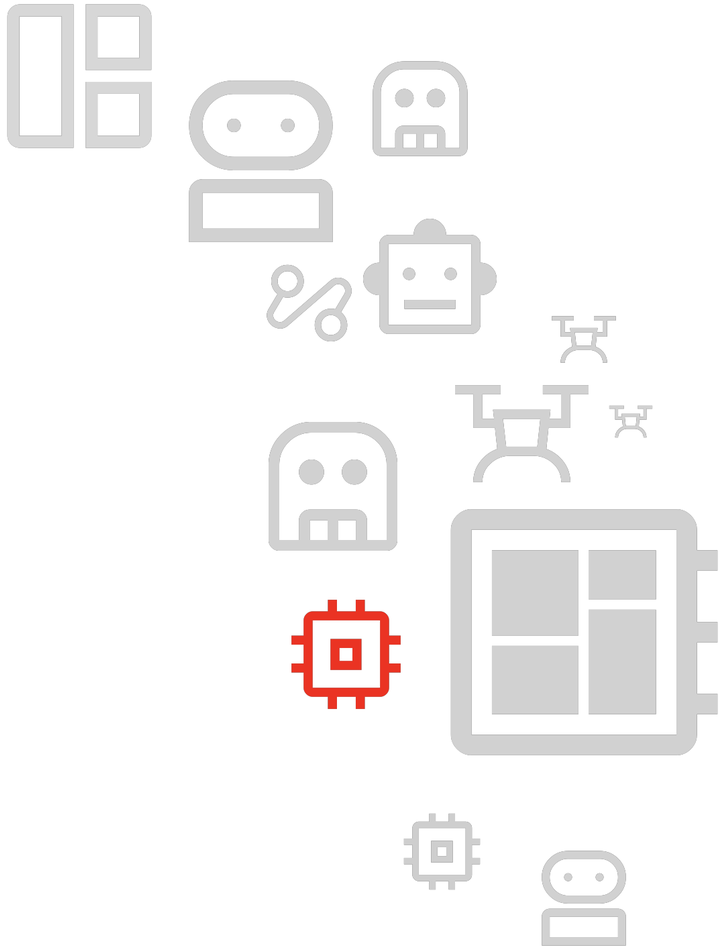
UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Soft actor-critic

Variação do método Actor-Critic que busca maximizar a combinação de recompensa esperada e entropia da política.

- Algoritmo de entropia máxima
- Espaço de ação contínuo
- Off-policy
- Model-free
- Estado-da-arte para tarefas de robótica com controle contínuo



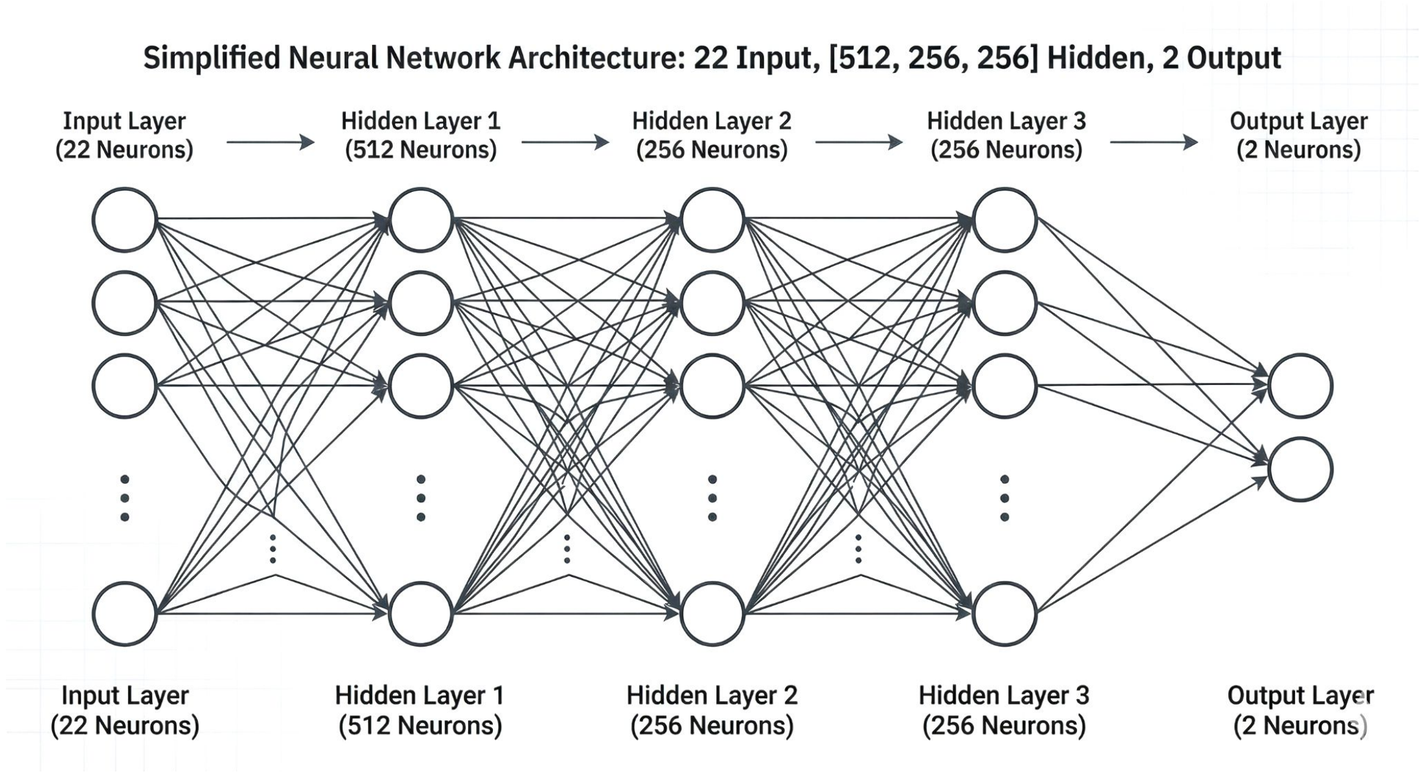
UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Configuração Paramétrica

Arquitetura de rede neural

Definimos a mesma arquitetura de rede para o ator e o crítico, consistindo em uma rede de três camadas ocultas de 512, 256 e 256 neurônios, respectivamente.



UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS

Outros hiperparâmetros

- **Taxa de aprendizado:** 0.0003
- **Tamanho do buffer:** 1.000.000
- **Tamanho do minibatch:** 512
- **gamma:** 0.99
- **tau:** 0.005

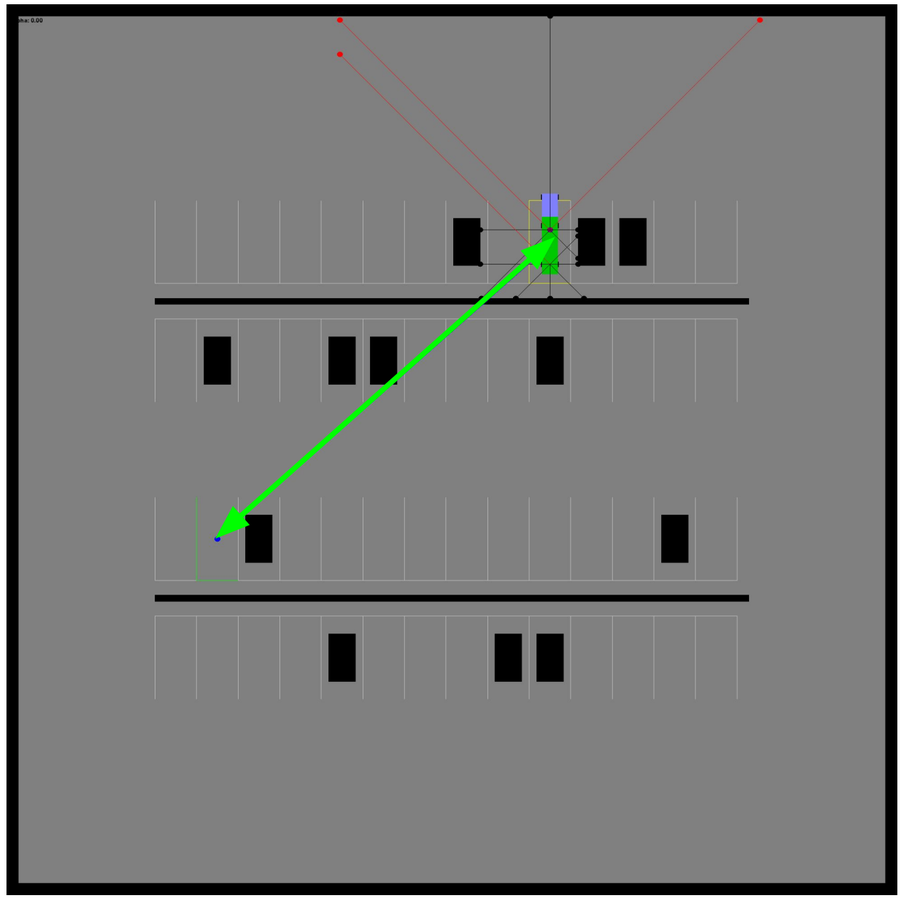
Heurísticas

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Distância L2 (euclidiana)

Usamos a distância euclidiana do veículo até a vaga destino como sinal de recompensa. O agente é recompensado por reduzir essa distância a cada passo de simulação.

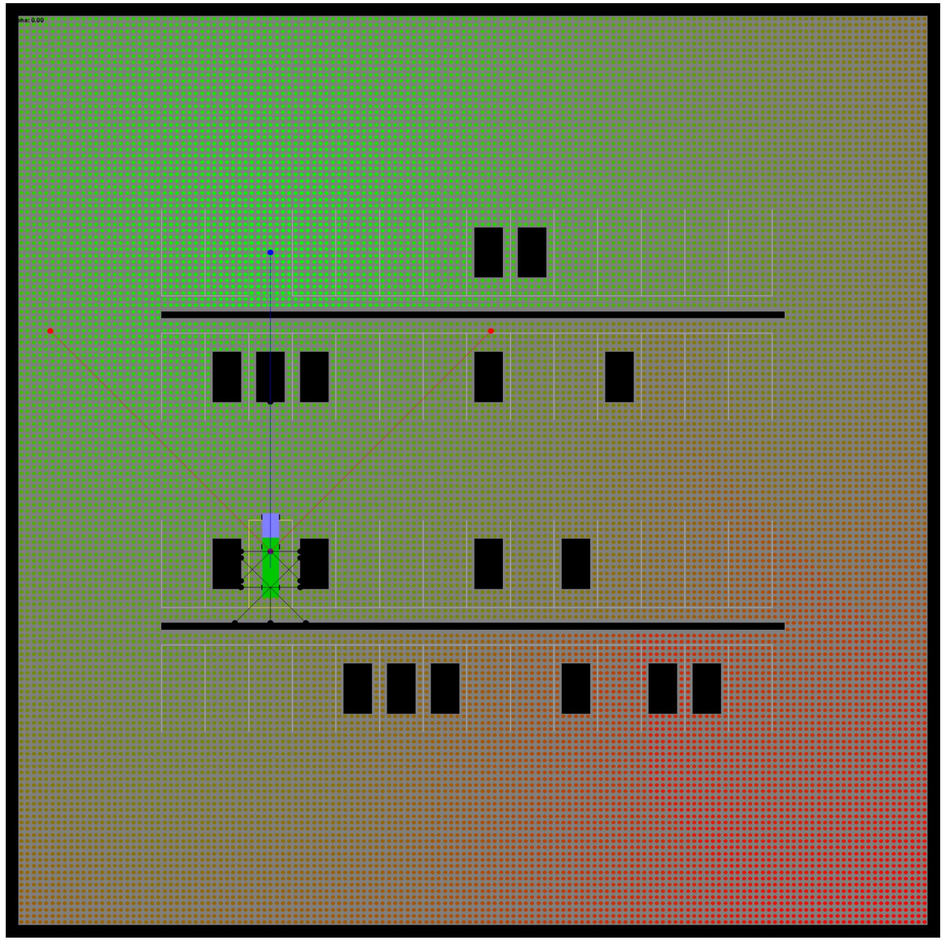


UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

BFS

Primeiramente, discretizamos o mapa em pixels de 1.0m² e usamos o BFS (Breadth-first-search) a partir do alvo para computar um **mapa de distâncias** considerando todos os obstáculos do ambiente. Então, a recompensa do agente é calculada consultando-se esse mapa a cada passo e computando a diferença entre a distância anterior - também pelo BFS - e a atual.



UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

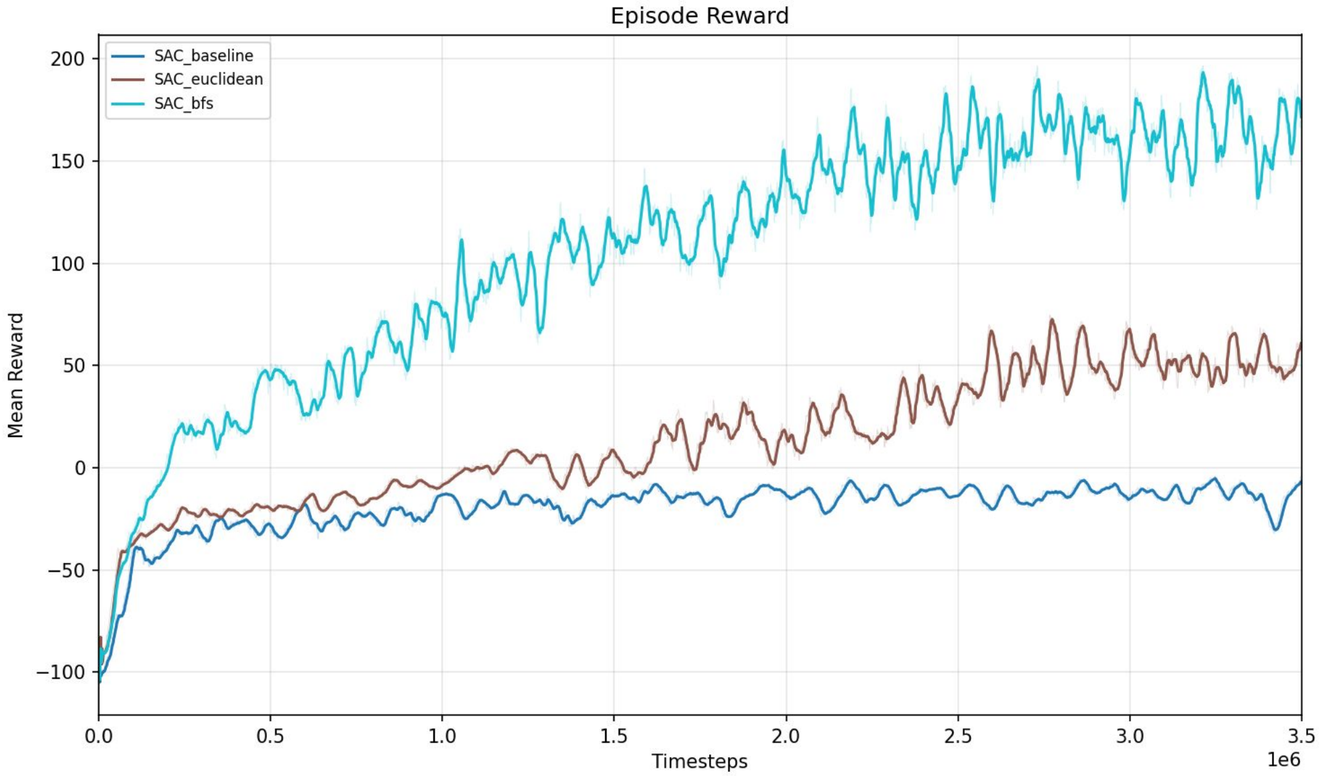
.inf
INSTITUTO DE INFORMÁTICA
UFRGS

Resultados

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Recompensa Média

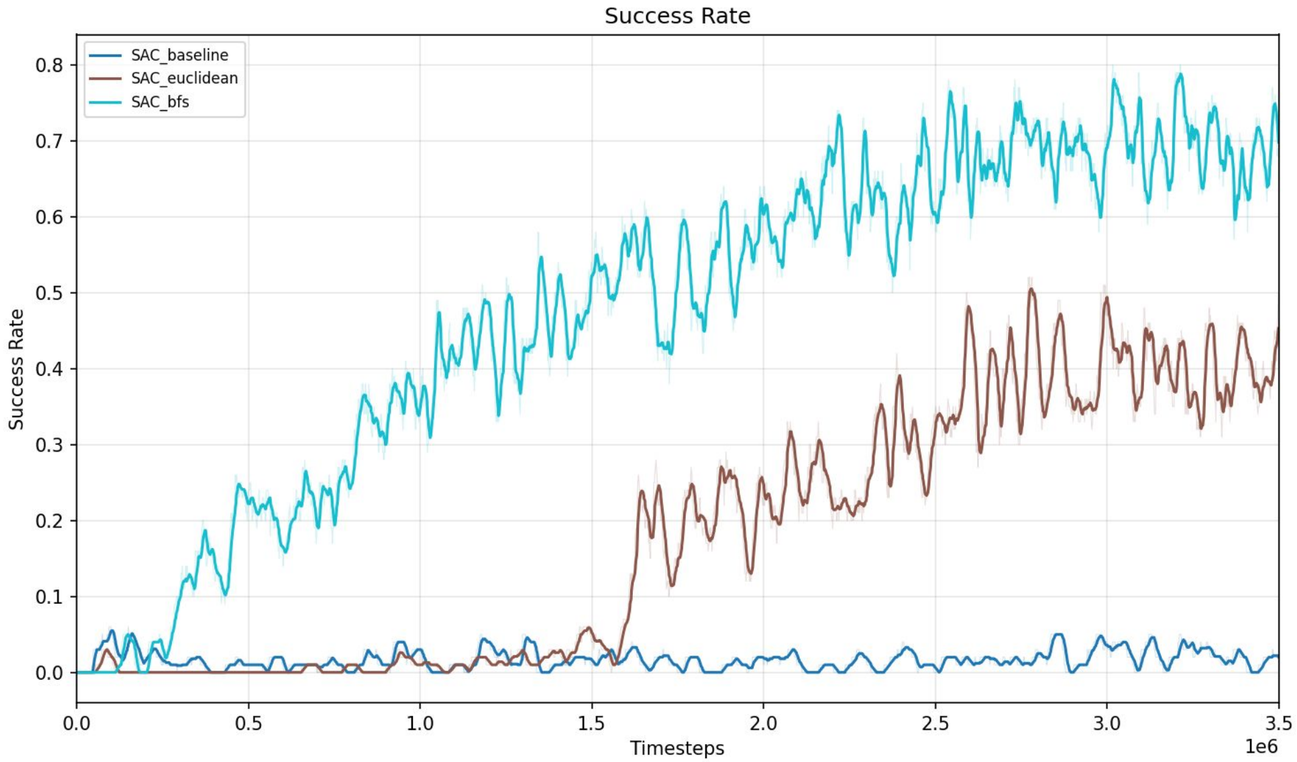


- Vantagem do BFS para o reward shaping fica evidente.
- O modelo que usa a heurística euclidiana tardou em encontrar o objetivo, e estabilizou a recompensa média em um limiar mais baixo.
- Baseline estabilizou dentro da faixa de recompensa negativa, indicando que ele aprendeu a 'sobreviver' (sem colisões), mas sem chegar ao objetivo

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Taxa de Sucesso

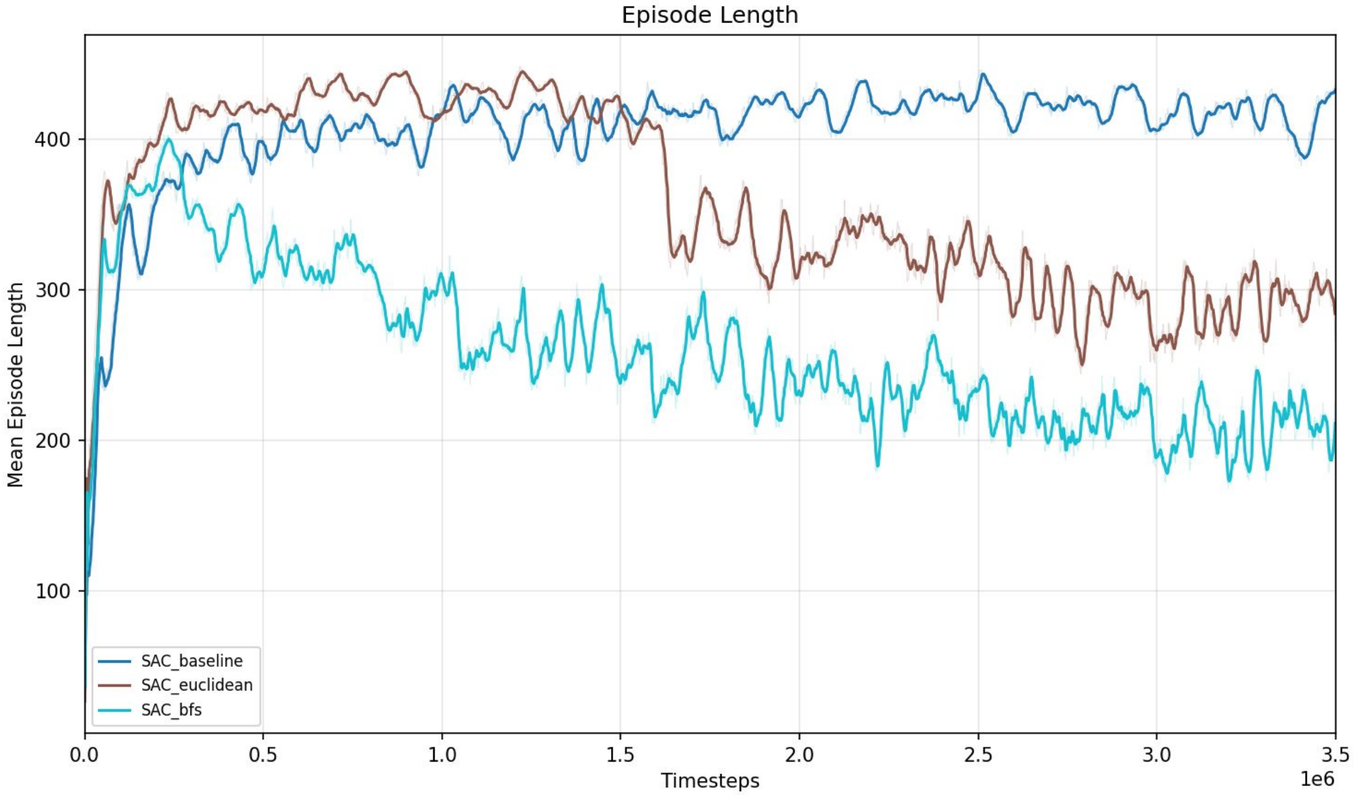


- Definimos taxa de sucesso como a frequência que o agente atinge o objetivo principal - estacionar na vaga.
- Novamente, BFS se sobressai com taxa de sucesso significativamente maior
- Heurística euclidiana estabiliza em um patamar menor - Obtém sucesso nos casos que não há obstáculo entre início e objetivo

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Duração de episódio



- a heurística mais 'inteligente' não apenas elevou a taxa de sucesso, mas também a velocidade de resolução do objetivo

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Conclusões

Potencial de Reward shaping com HuRL

O HuRL apresenta um grande potencial de aumentar a eficiência do treinamento de agentes de RL em aplicações de planejamento de rotas. Em nosso caso, o aprendizado efetivo só foi possibilitado pelo uso de uma heurística.

- O reward shaping ajuda a superar a esparsidade recompensa do problema de estacionamento de veículos

Dependência da heurística

O vantagem trazida pelo reward shaping depende fortemente da heurística escolhida. Quanto mais a heurística se conforma à dinâmica real do ambiente, maior o ganho de eficiência no treinamento do agente.

Trabalhos futuros

Avaliação com outras heurísticas

Analisar o desempenho de outras heurísticas, como Dijkstra, A*, etc, incluindo heurísticas não-holonômicas.

Uso de Coordenadas Privilegiadas

No decorrer do desenvolvimento, tomamos conhecimento das coordenadas privilegiadas, que podem trazer informações diretivas para o modelo que tem potência de levar a um aprendizado mais rápido, ficando como trabalho futuro.

Exploração de outros algoritmos de RL

Analisar o desempenho de outros algoritmos de controle contínuo off-policy, como TD3, ou on-policy, como o PPO sob a ótica do HuRL, se faz necessário. Diferentes algoritmos de RL podem se beneficiar do reward shaping em magnitudes distintas.

Referências

References

[1] Ching-An Cheng, Andrey Kolobov, and Adith Swaminathan. Heuristic-guided reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13550–13563, 2021.

[2] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[4] Chenyong Guan and Yu Jiang. A tractor-trailer parking control scheme using adaptive dynamic programming. *Complex & Intelligent Systems*, 8(3):1835–1845, 2022.

[5] Mario Rosenfelder, Henrik Ebel, Jasmin Krauspenhaar, and Peter Eberhard. Model predictive control of non-holonomic systems: Beyond differential-drive vehicles. *Automatica*, 152:110972, 2023.

Obrigado!

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS