Aprendendo a Andar na Linha

Aprendizado de Máquina por Reforço aplicado à locomoção de quadrúpedes para estabilização vertical

André Luís A. de Souza
Instituto de Informática
UFG
Goiânia, Goiás

Luan Gabriel S. Oliveira
Instituto de Informática
UFG
Goiânia, Goiás

Marcelo Henrique L. Ferreira
Instituto de Informática
UFG
Goiânia, Goiás

Pedro M. Bittencourt
Instituto de Informática
UFG
Goiânia, Goiás

***Abstract*—Esse artigo apresenta um estudo aplicando Aprendizado de Máquina por Reforço na locomoção de robôs quadrúpedes em um ambiente simulado altamente paralelizável. A inspiração para o estudo se dá a partir da necessidade de um controle mais estável para aplicações robóticas no meio agrícola, como coleta de dados e tarefas de percepção.**

***Index Terms*—*Reinforcement Learning*, robótica, *Agrotech*, locomoção, controle.**

I. Introdução e revisão bibliográfica

O controle de caminhada de robôs quadrúpedes em ambientes agrícolas apresenta desafios significativos, especialmente em termos de odometria e coleta de dados. O ambiente irregular e variado contribui para a instabilidade no eixo Z, dificultando a precisão dos algoritmos de odometria e SLAM (Simultaneous Localization and Mapping). No laboratório do Pequi Mecânico, experimentos [1] indicaram que essa instabilidade pode estar comprometendo o desempenho desses algoritmos. Portanto, o objetivo deste estudo é investigar se modificações na função de recompensa, focando na estabilidade do eixo Z, podem melhorar o desempenho desses algoritmos.

Literatura A base teórica deste estudo é fundamentada em diversas fontes:

- **"Reinforcement Learning: An Introduction" de Sutton e Barto [2]**: Este livro oferece uma visão abrangente sobre o aprendizado por reforço (RL), que é o alicerce teórico do nosso estudo.
- **Walk These Ways [4]**: Este trabalho explora o uso de RL para controle de caminhada em robôs, apresentando métodos e resultados relevantes para nossa pesquisa.
- **Legged Gym [3]**: Um repositório que utiliza a plataforma IsaacGym para simulação e treinamento de robôs quadrúpedes com RL, especificamente utilizando o algoritmo PPO (Proximal Policy Optimization).
- **IsaacGym [6]**: Uma plataforma de simulação paralelizada desenvolvida pela NVIDIA, ideal para treinamento de algoritmos de RL em ambientes simulados de alta fidelidade.
- **PPO (Proximal Policy Optimization) [5]**: Um algoritmo de RL amplamente utilizado devido à sua eficiência e estabilidade em diversas tarefas de controle, incluindo a caminhada de robôs.

Os dados utilizados para os experimentos foram gerados através de simulações no IsaacGym, seguindo o modelo do Legged Gym. Essas simulações fornecem dados detalhados sobre a dinâmica do robô, incluindo a variação no eixo Z durante a caminhada.

Para avaliar o impacto das modificações na função de recompensa, utilizamos o algoritmo PPO. Modificamos os pesos das recompensas relacionadas à estabilidade do eixo Z e comparamos os resultados com a baseline fornecida pelo Legged Gym.

A avaliação dos resultados foi feita através da análise da variação no eixo Z durante as simulações de teste. As métricas de desempenho incluem a variância do eixo Z e a estabilidade geral do robô durante a caminhada. Devido à ausência de um robô físico, a análise foi limitada às simulações.

II. Fundamentos teóricos

A. Algoritmo PPO

O PPO é um algoritmo de aprendizado por reforço que busca otimizar a política de um agente de maneira estável e eficiente. Ele utiliza uma abordagem de "clipping" para garantir que as atualizações de política não sejam muito grandes, o que pode levar à instabilidade. A função de recompensa no PPO é crucial para direcionar o comportamento do agente, e neste estudo, focamos em modificar essa função para melhorar a estabilidade do eixo Z em robôs quadrúpedes.

B. Legged Gym

Legged Gym é uma implementação que utiliza IsaacGym para simular e treinar robôs quadrúpedes com o PPO. Ele fornece uma base sólida para desenvolver e testar controladores de caminhada. A plataforma permite a paralelização das simulações, acelerando o processo de treinamento e avaliação dos algoritmos de RL. Utilizamos como base o artigo do legged gym, que é um artigo/repositório que serve como uma boa base para criar uma simulação de quadrúpedes no isaacgym. Esse repositório utiliza o PPO. A ideia é mexer nas funções de recompensa presentes nesse algoritmo.

III. Metodologia

O primeiro passo da nossa metodologia foi um estudo aprofundado da API do IsaacGym, uma plataforma de simulação desenvolvida pela NVIDIA que permite a criação de ambientes virtuais de alta fidelidade para o treinamento de robôs. Esse estudo envolveu a leitura detalhada da documentação

oficial, análise de exemplos fornecidos pela NVIDIA e a implementação de scripts básicos para entender a configuração dos ambientes e o controle dos robôs. Exploramos também as capacidades de paralelização da plataforma para otimizar o tempo de treinamento dos algoritmos de aprendizado por reforço, o que é essencial para lidar com a complexidade das simulações de robôs quadrúpedes.

Em seguida, reproduzimos os experimentos do Legged Gym para estabelecer uma baseline de desempenho. Para isso, configuramos o ambiente de simulação do Legged Gym no Isaac-Gym, garantindo que todos os parâmetros, como o terreno, a dinâmica do robô e as condições de simulação, estivessem corretos. Executamos o treinamento do algoritmo PPO com a configuração padrão do Legged Gym, utilizando múltiplos processos paralelos para acelerar a convergência. Durante este treinamento, monitoramos e registramos as métricas de desempenho, incluindo a variação do eixo Z, tanto durante o treinamento quanto nas simulações de teste, o que nos permitiu obter uma linha de base robusta para comparação.

A função de recompensa do PPO, crucial para direcionar o comportamento do robô, foi então modificada para melhorar a estabilidade no eixo Z. Estudamos a função de recompensa original do Legged Gym para identificar os componentes que impactam diretamente a estabilidade vertical do robô. A partir dessa análise, ajustamos os pesos desses componentes para incentivar a minimização da variação no eixo Z, aumentando o peso da penalização para grandes desvios verticais e ajustando outros parâmetros para manter um equilíbrio no desempenho geral do robô. Essas modificações foram projetadas para reduzir a instabilidade no eixo Z, que é um problema significativo em ambientes agrícolas.

Finalmente, realizamos um novo ciclo de treinamento utilizando a função de recompensa modificada, mantendo a mesma configuração de ambiente da baseline para garantir uma comparação justa. Monitoramos continuamente as métricas de desempenho durante o treinamento para verificar se as modificações estavam produzindo os resultados esperados. Para avaliar a eficácia das modificações, realizamos simulações de teste com ambos os modelos, comparando a variância do eixo Z. Utilizamos gráficos e tabelas para visualizar as diferenças, o que nos permitiu concluir que as modificações na função de recompensa resultaram em uma redução significativa na variação do eixo Z, indicando uma melhoria na estabilidade do robô quadrúpede.

IV. Resultados e Conclusões

Os resultados obtidos com as simulações demonstraram uma melhora sutil na variância da velocidade no eixo Z do robô quadrúpede. A seguir, apresentamos os dados em gráficos e tabelas que ilustram a comparação entre a baseline e o modelo modificado.

Reward	Avarage
Velocidade Angular XY	-0.008
Velocidade Linear Z	-0.012

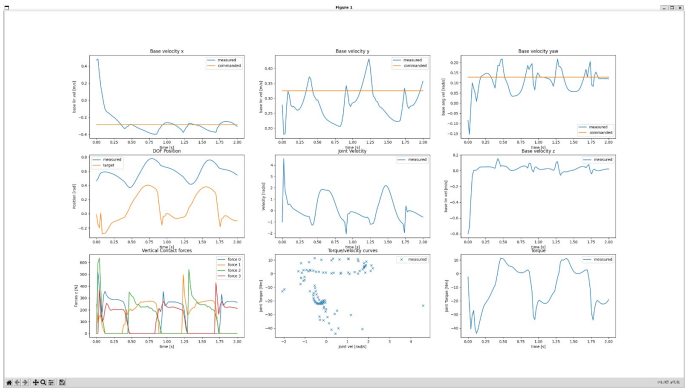

Fig. 1. Variação da velocidade no eixo Z na baseline

Fig. 2. Variação da velocidade no eixo Z no modelo modificado

Reward	Avarage
Velocidade Angular XY	-0.006
Velocidade Linear Z	-0.007

Os gráficos mostram a diferença na variação da velocidade no eixo Z entre a baseline e o modelo modificado, destacando a redução na variância. A tabela compara os valores numéricos da variância e da média da velocidade, evidenciando que, embora a melhora seja sutil, ela é consistente com as modificações realizadas na função de recompensa.

A análise qualitativa desses resultados indica que a modificação na função de recompensa contribuiu para uma maior estabilidade no eixo Z, o que pode ter implicações positivas para o desempenho dos algoritmos de odometria e SLAM. No entanto, é fundamental validar esses resultados em um robô físico. A próxima etapa ideal seria implementar o algoritmo treinado em um robô real e avaliar se a maior estabilidade no eixo Z resulta em um desempenho superior dos algoritmos de odometria em ambientes reais.

Os experimentos realizados mostraram que é possível reduzir a variação da velocidade no eixo Z de robôs quadrúpedes através de ajustes na função de recompensa do PPO. Embora a melhoria observada tenha sido sutil, ela aponta para a viabilidade da abordagem. Este estudo serviu como uma fase preliminar para validar a modificação, e os conhecimentos

adquiridos serão fundamentais para os futuros trabalhos no laboratório do Pequi Mecânico.

Os próximos passos incluem a implementação do algoritmo modificado em um robô físico para validar os resultados observados nas simulações. Além disso, exploraremos outras funções de recompensa e algoritmos de RL que possam oferecer melhorias adicionais na estabilidade e desempenho dos robôs quadrúpedes em ambientes reais.

REFERENCES

[1] bitdog - Google Drive. Disponível em: ¡https://drive.google.com/drive/folders/1f6DDEzuSrpHcGhzv8kB2pK39HUe0C4QW¿. Acesso em: 11 jul. 2024.

[2] SUTTON, R. S.; BARTO, A. Reinforcement learning : an introduction. Cambridge, Ma ; Lodon: The Mit Press, 2018.

[3] RUDIN, N. et al. Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. Disponível em: ¡https://arxiv.org/abs/2109.11978¿.

[4] MARGOLIS, G. B.; AGRAWAL, P. Walk These Ways: Tuning Robot Control for Generalization with Multiplicity of Behavior. arXiv (Cornell University), 1 jan. 2022.

[5] SCHULMAN, J. et al. Proximal Policy Optimization Algorithms. Disponível em: ¡http://arxiv.org/abs/1707.06347¿.

[6] Isaac Gym - Preview Release. Disponível em: ¡https://developer.nvidia.com/isaac-gym¿.