Deep Reinforcement Learning para Otimização de Trajetórias de UAVs em Cenários de Desastre Utilizando DDPG em Ambientes Contínuos.

João Paulo de Souza & Raul Moreno Pereira

I. Relevância e Motivação

A ocorrência de desastres naturais no Brasil, como enchentes e ventos fortes, tem se intensificado nos últimos anos em função das mudanças climáticas e urbanização desordenada. Esses eventos, além de provocaram perdas humanas e materiais, frequentemente comprometem a infraestrutura de comunicação terrestre, dificultando a coordenação das equipas de resgate e a comunicação entre as vítimas. O uso de veículos aérios não tripulados (UAVs, do inglês *Unmanned Aerial Vehicles*) como estações base móveis desponta como uma solução tecnológica de alto impacto, visto que sua mobilidade, baixo custo de operação e capacidade de rápida implementação permitem restabelecer a conectividade em áreas críticas de difícil acesso.

Embora soluções já tenham explorado a utilização de UAVs em tais cenários, muitas delas se limitam ao emprego de métodos de otimização clássicos ou de algoritmos de aprendizado por reforço em sua forma tabular. O trabalho de Zhao et al. (2022) [1] mostrou que a aplicação de Q-Learning pode auxiliar na determinação de posições mais eficientes para UAVs, garantindo cobertura a usuários desassistidos. Entretanto, o método tabular apresenta limitações importantes: exige discretização do espaço de estados e ações, reduzindo sua capacidade de adaptação a cenários contínuos e tridimensionais, além de aumentar o tempo de convergência do aprendizado. Além disso, o Brasil possui particularidades geográficas e climáticas que não podem ser ignoradas. Situações como enchentes urbanas, comuns em grandes centros, e rajadas de vento que afetam a estabilidade de voo dos drones em áreas costeiras ou do interior, exijam modelos de simulação mais realistas pata o treinamento dos agentes. Nesse contexto, torna-se relevante investigar o emprego de algoritmos de aprendizado de reforço profundo, como o *Deep Deterministic Policy Gradient* (DDPG), que oferecem mecanismos mais robustos para lidar com espaços de ação contínuos.

Além disso, a utilização do Unity como plataforma de simulação acrescenta uma dimensão inovadora ao estudo, visto que permite a criação de ambientes tridimensionais realistas, incorporando variáveis como, principalmente, efeitos climáticos como chuvas intensas, rajadas de vento e inundações, sendo utilizado também como sistema de recompensas, permitindo avaliar em tempo real a estabilidade do voo, o consumo energético e a cobertura de usuários, fornecendo feedback direto para o processo de aprendizado de reforço.

II. Revisão da Literatura

A literatura científica acerca da utilização de UAVs em redes de comunicação emergências tem crescido de forma significativa, acompanhando os avanços tecnológicos em inteligência artificial e nas redes móveis de última geração. Em um dos trabalhos pioneiros sobre o tema, Lyu et al. (2016) [2] propuseram a otimização da colocação de UAVs para cobertura máxima em cenários de comunicação móvel, destacando os desafios relacionados à limitação de energia e à cobertura limitada de cada drone. Posteriormente, Mozaffari et al. (2019) [3] realizaram uma revisão abrangente sobre o uso de UAVs em redes sem fio, discutindo aplicações, desafios técnicos e lacunas de pesquisa, ressaltando a importância da trajetória ótima e do uso de técnicas de aprendizado inteligente para tomada de decisão em tempo real.

No contexto específico de cenários de desastre, Zhao et al. (2022) [1] apresentam um modelo baseado em Q-Learning tabular para otimizar o posicionamento de UAVs com cooperação com estações base terrestres parcialmente funcionais, buscando aumentar a cobertura de usuários não atendidos. Embora os resultados tenham demostrado a viabilidade da abordagem, as limitações do Q-Learning tabular, sobretudo a necessidade de discretizar o espaço contínuo de estados e ações, reduziram a aplicabilidade prática do método em cenários complexos e dinâmicos. Essas limitações têm motivado a comunidade científica a migrar para algoritmos de aprendizado por reforço profundo.

Entre os algoritmos mais explorados, destacam-se o Deep Q-Network (DQN), introduzido por Mnih et al. (2015) [4], que foi um marco ao demonstrar a capacidade de redes neurais profundas em aproximar funções de valor em ambientes de alta dimensionalidade. Contudo, o DQN permanece restrito a espaços de ação discretos, o que limita sua aplicação em UAVs que operam em ambientes contínuos. Nesse sentido, o DDPG, proposto por Lillicrap et al. (2015) [5], surge como alternativa robusta ao estender o aprendizado profundo para espaços de ação contínuos, permitindo ajustes mais refinados em variáveis como posição, velocidade e altitude.

Aplicações de aprendizado por reforço profunso em UAVs vêm se multiplicando. Liu, Liu e Chen (2019) [6] exploraram a utilização de aprendizado por reforço em redes com múltiplos UAVs para otimizar cobertura e movimentação, evidenciando os benefícios do uso de algoritmos avançados em cenários dinâmicos. Zeng e Zhang (2017) [7] também destacaram a importância da otimização de trajetória em UAVs para

melhorar a eficiência energética das redes sem fio, apontando que métodos tradicionais de otimização matemática se tornam rapidamente inviáveis à medida que a complexidade do ambiente cresce.

III. Problema a ser Resolvido

O problema central deste projeto consiste em desenvolver uma abordagem eficaz para a otimização das trajetórias de UAVs em cenários pós-desastre no Brasil, de modo a maximizar a cobertura de usuários não atendidos, minimizar o tempo de convergência do aprendizado e assegurar uma comunicação estável em ambientes complexos. A questão que se coloca é como substituir o Q-Learning tabular por um algoritmo capaz de explorar o espaço contínuo tridimensional em que os UAVs operam.

Em ambientes reais, os UAVs precisam realizar movimentos contínuos em três dimensões, ajustando altitude, velocidade e direção em tempo real, ao mesmo tempo em que consideram restrições como energia limitada, interferência de sinal e obstáculos físicos. Nesse contexto, os métodos tradicionais de aprendizado tabular tornam-se insuficientes, tornando importante o uso de técnicas mais avançadas que consigam aprender políticas de controle diretamente em espaços contínuos.

IV. Hipótese para a resolução do Problema

A hipótese norteadora deste projeto é que a utilização do DDPG em conjunto com o Unity, atuando como sistema de recompensas, permitirá que os UAVs aprendam políticas de voo mais eficientes, resilientes e adaptativas do que as obtidas com Q-Learning tabular. Acredita-se que, ao simular cenários climáticos diversos, será possível treinar os UAVs em condições próximas às encontradas no Brasil, possibilitando maior cobertura de usuários, menor tempo de convergência do aprendizado e maior eficiência energética, mesmo em cenários adversos. Dessa forma, espera-se que o modelo proposto não apenas supere abordagens anteriores, mas também ofereça soluções mais realistas e aplicáveis.

V. Descrição da abordagem

A abordagem proposta será implementada a partir da integração entre o Unity e o algoritmo de aprendizado por reforço profundo DDPG. O Unity desempenhará papel central não apenas como ambiente tridimensional de simulação, mas também como sistema de geração de recompensas. Para isso, serão criados cenários urbanos representativos de cidades brasileiras frequentemente atingidas por enchentes e ventos fortes, incorporando elementos dinâmicos, como áreas inundadas, obstáculos em movimento e rajadas de vento que interferem na trajetória dos UAVs.

Nesse ambiente, os UAVs serão modelados como agentes inteligentes dotados de sensores virtuais que lhes permitem coletar informações sobre posição, energia remanescente, intensidade das forças climáticas, qualidade do sinal e número de usuários atendidos. Esses dados serão processados pelo sistema de aprendizado, enquanto o Unity calculará em tempo real os efeitos das ações tomadas, retornando valores de recompensa ao agente. Essa recompensa será definida por uma função que considera positivamente a ampliação da cobertura de usuários, a estabilidade de voo e a eficiência energética, ao mesmo tempo em que aplica penalização para colisões, trajetórias ineficientes, consumo excessivo de bateria e perda de conectividade devido às condições climáticas adversas. O algoritmo DDPG será responsável por atualizar as políticas determinísticas de controle contínuo com base nos sinais de recompensa fornecidos pelo Unity. Isso permitirá que os UAVs aprendam trajetórias adaptativas em três dimensões, ajustando altitude, velocidade e direção de forma precisa mesmo em ambientes instáveis. A comunicação entre o Unity e o DDPG serpa medida por interfaces como Unity ML-Agents, que possibilitam a troca contínua de estados, ações e recompensas.

Fig. 1. Inicio da construção da simulação na plataforma Unity.

Serão também conduzidos experimentos comparativos entre a proposta baseada em DDPG e métodos tabulares como o Q-Learning. Serão analisadas métricas como o tempo de convergência, a qualidade de usuários atendidos, a eficiência energética e a resiliência frente a condições adversas simuladas no Unity. Essa estratégia permitirá validar a hipótese de que a utilização do Unity como sistema de recompensa, aliado ao DDPG, resulta em um modelo mais robusto, eficiente e aplicável aos cenários desastre recorrentes no Brasil.

VI. Contribuições Esperadas

As contribuições deste projeto podem ser divididas em três dimensões. No âmbito científico, espera-se demonstrar a superioridade do DDPG em relação ao Q-Learning em ambientes contínuos e adversos, reforçando o papel dos algoritmos de aprendizado por reforço profundo na otimização de UAVs. No âmbito tecnológico, a utilização do Unity como sistema de recompensas, capaz de simular enchentes e ventos fortes, representa uma inovação ao aproximar a simulação acadêmica das condições reais enfrentadas no Brasil. Por fim, no âmbito social, este projeto pode subsidiar a criação de ferramentas de suporte à comunicação emergencial em situações críticas, reduzindo o impacto de desastres naturais sobre comunidades vulneráveis.

Referências

[1] Shiye Zhao, Kaoru Ota e Mianxiong Dong. *UAV Base Station Trajectory Optimization Based on Reinforcement Learning in Post-disaster Search and Rescue Operations*. 2022. arXiv: 2202.10338 [cs.LG]. URL: https://arxiv.org/abs/2202.10338.

[2] Jiangbin Lyu, Yong Zeng, Rui Zhang e Teng Joon Lim. "Placement Optimization of UAV-Mounted Mobile Base Stations". Em: *IEEE Communications Letters* 21.3 (2017), pp. 604–607. DOI: 10.1109/LCOMM.2016.2633248.

[3] Mohammad Mozaffari, Walid Saad, Mehdi Bennis, Young-Han Nam e Mérouane Debbah. "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems". Em: *IEEE Communications Surveys Tutorials* 21.3 (2019), pp. 2334–2360. DOI: 10.1109/COMST.2019.2902862.

[4] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". Em: *Nature* 518 (2015), pp. 529–533. URL: https://api.semanticscholar.org/CorpusID:205242740.

[5] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver e Daan Wierstra. "Continuous control with deep reinforcement learning". Em: *CoRR* abs/1509.02971 (2015). URL: https://api.semanticscholar.org/CorpusID:16326763.

[6] Xiao Liu, Yuanwei Liu, Yue Chen, Luhan Wang e Zhaoming Lu. "Machine Learning Aided Trajectory Design and Power Control of Multi-UAV". Em: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9014216.

[7] Yong Zeng e Rui Zhang. "Energy-Efficient UAV Communication With Trajectory Optimization". Em: *IEEE Transactions on Wireless Communications* 16.6 (2017), pp. 3747–3760. DOI: 10.1109/TWC.2017.2688328.