Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Computação Inteligente ou Modelagem Computacional

Título: Desvendando os Segredos da IA em Jogos: Explicações Contrafactuais com Insights Introspectivos para Agentes de RL Mais Inteligentes e Transparentes

Orientador(a): Pablo Barros (pablo.barros@sony.com)

Este projeto de pesquisa visa aprimorar agentes de aprendizado por reforço (RL) em ambientes de jogos, incorporando explicações contrafactuais na estrutura de recompensas do agente. O objetivo é não apenas melhorar o desempenho dos agentes baseados em RL, mas também fornecer explicações claras e confiáveis para o processo de tomada de decisão do agente. Essa abordagem pretende facilitar o surgimento de estratégias distintas durante o treinamento, oferecendo insights mais profundos sobre como e por que ações específicas são escolhidas pelo agente. A integração de explicações contrafactuais pode aumentar significativamente a transparência e a interpretabilidade dos agentes de RL, atendendo a uma necessidade crítica no desenvolvimento de sistemas inteligentes de jogos.

O problema central abordado nesta pesquisa é a falta de interpretabilidade nos agentes baseados em RL atuais para jogos competitivos. Embora esses agentes possam alcançar altos níveis de desempenho, entender a lógica por trás de suas decisões permanece um desafio [1]. Essa opacidade pode prejudicar a confiança e limitar a usabilidade dos agentes de RL em aplicações complexas. A literatura existente fornece vários métodos para melhorar a explicabilidade, como o uso de mecanismos de atenção [2] e a geração de explicações em linguagem natural [3]. No entanto, essas abordagens muitas vezes não se relacionam diretamente com as estruturas de recompensas que orientam o processo de aprendizado do agente, levando a uma lacuna na compreensão e otimização abrangente do comportamento do agente.

A solução proposta envolve a extensão do conceito de medições introspectivas — avaliações de probabilidade das ações que levam às recompensas máximas com base nos valores-Q do agente — para aprimorar a explicabilidade dos grandes modelos de linguagem (LLMs) usados em agentes de RL para jogos. Ao integrar essas medições introspectivas com explicações contrafactuais, o projeto busca criar um processo de aprendizado mais transparente. Essa abordagem pode aproveitar estruturas existentes em raciocínio contrafactual e IA explicável, fornecendo um método inovador para gerar insights acionáveis e estratégias distintas. Os resultados esperados incluem desempenho aprimorado dos agentes, processos de tomada de decisão mais robustos e explicáveis, e o potencial para estabelecer novos padrões no desenvolvimento de sistemas inteligentes de jogos.

Referências Bibliográficas:

[1] Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, *214*, 106685.

[2] Wang, X., Chen, Y., Yang, J., Wu, L., Wu, Z., & Xie, X. (2018, November). A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)* (pp. 587-596). IEEE.

[3] Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Liu, T., ... & Liu, N. (2024). Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint arXiv:2403.08946*.