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

Proposta de Tese de Doutorado

Área: Inteligência Computacional / Etica e inteligencia artificial

Título: Desaprendiagem de maquina: Ensinando grandes modelos fundamentais de linguagem e visão a esquecer o que é (eticamente) errado

Orientador – Pablo Vinicius Alves de Barros (pablo.barros@sony.com)

Coorientador – Bruno José Torres Fernandes (bjtf@ecomp.poli.br)

Descrição

O aprendizado de máquina é amplamente utilizado para melhorar a capacidade de sistemas em realizar tarefas específicas através da análise de dados, mas a necessidade de desenvolver técnicas para reverter ou desfazer esse aprendizado está crescendo. O campo emergente do "machine unlearning" [1] busca criar métodos para que as máquinas "esqueçam" informações indesejáveis ou obsoletas, aumentando a flexibilidade e adaptabilidade dos sistemas de inteligência artificial. Este projeto visa investigar algoritmos e técnicas para identificar, remover e reverter conhecimento previamente adquirido, além de explorar as implicações éticas e de privacidade relacionadas [2].

O projeto envolverá a investigação de algoritmos, técnicas e abordagens para identificar, remover e reverter conhecimento previamente adquirido em modelos de aprendizado de máquina. Além disso, serão exploradas implicações éticas e questões de privacidade relacionadas ao desaprendizado de máquina [3]. Espera-se que os resultados deste projeto contribuam para o desenvolvimento de estratégias eficazes de desaprendizado de máquina e abram novas possibilidades para a aplicação responsável e ética da inteligência artificial.

O candidato ideal possui conhecimento em aprendizado de máquina, incluindo fundamentos de estatística, probabilidade e inteligência artificial, além de experiência prática no desenvolvimento de redes neurais artificiais. Proeficiência na língua inglesa é essencial para colaborar com grupos de pesquisa internacionais. Espera-se que os resultados deste projeto contribuam significativamente para o desenvolvimento de estratégias eficazes de desaprendizado de máquina, abrindo novas possibilidades para a aplicação responsável e ética da inteligência artificial.

Referências Bibliográficas

1. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021, May). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 141-159). IEEE.
2. Gupta, V., Jung, C., Neel, S., Roth, A., Sharifi-Malvajerdi, S., & Waites, C. (2021). Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, *34*, 16319-16330.
3. Zhang, D., Pan, S., Hoang, T., Xing, Z., Staples, M., Xu, X., ... & Zhu, L. (2023). To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *arXiv preprint arXiv:2302.03350*.