Título: Metodologias de Otimização e Robustez para Modelos de Aprendizagem de Máquina Eficientes em bordas (Edge-AI)

Orientador: Erick de Andrade Barboza

Resumo:

O avanço recente dos modelos de Aprendizagem de Máquina (AM), especialmente os Grandes Modelos de Linguagem (*Large Language Models* - LLMs) e seus equivalentes compactos (*Small Language Models* - SLMs), transformou radicalmente o panorama da Inteligência Artificial. No entanto, a dependência severa de infraestruturas de computação em nuvem de alto desempenho impõe gargalos críticos de latência, privacidade, consumo energético e largura de banda.

A migração desses modelos para a borda da rede (*Edge Computing*) e sistemas embarcados com recursos restritos é o próximo grande passo científico e tecnológico. Contudo, essa transição exige viabilizar a execução eficiente sem degradar severamente a acurácia dos modelos, além de garantir a segurança dessas aplicações em cenários do mundo real.

O objetivo geral deste projeto é investigar, propor e desenvolver metodologias, algoritmos e frameworks para a criação de modelos de aprendizagem de máquina altamente eficientes, viabilizando a inferência e/ou o treino de alta performance diretamente em sistemas embarcados e dispositivos de borda.

De maneira específica, o projeto irá buscar avaliar e propor técnicas de compressão de modelos (como quantização extrema, poda de neurônios, destilação de conhecimento e *low-rank adaptation*) focadas na execução de LLMs e SLMs em hardware limitado. Além disso, outro objetivo específico será validar as metodologias desenvolvidas por meio de problemas do mundo real (ex: Internet das Coisas Médicas, cidades inteligentes, monitoramento industrial ou veículos autônomos). Por fim, também pretende-se investigar o impacto da compressão e da execução em hardware embarcado na resiliência dos modelos, avaliando explicitamente a degradação de desempenho frente a corrupções naturais nos dados (ruídos de sensores, falhas de captura) e a vulnerabilidade a ataques adversariais (*adversarial attacks*).

Referências Bibliográficas:

[1] MENGHANI, Gaurav. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. ACM Computing Surveys, v. 55, n. 12, p. 1-37, 2023.

[2] MURSHED, MG Sarwar et al. Machine learning at the network edge: A survey. ACM Computing Surveys (CSUR), v. 54, n. 8, p. 1-37, 2021.

[3] HONG, Junyuan et al. Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression. In: **International Conference on Machine Learning**. PMLR, 2024. p. 18611-18633.

[4] LIU, Zechun et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In: **Forty-first International Conference on Machine Learning**. 2024.

[5] WANG, Boxin et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. **Advances in Neural Information Processing Systems**, v. 36, p. 31232-31339, 2023.

[6] WANG, Fali et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. **ACM Transactions on Intelligent Systems and Technology**, v. 16, n. 6, p. 1-87, 2025.

[7] ZHENG, Yue et al. A review on edge large language models: Design, execution, and applications. **ACM Computing Surveys**, v. 57, n. 8, p. 1-35, 2025.

[8] HENDRYCKS, Dan; DIETTERICH, Thomas. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In: **International Conference on Learning Representations**, 2019.

[9] LIU, Chang et al. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. **International Journal of Computer Vision**, v. 133, n. 2, p. 567-589, 2025.

[10] CROCE, Francesco et al. RobustBench: a standardized adversarial robustness benchmark. In: **Conference on Neural Information Processing Systems**. 2021.

VAGAS: 2