Simpósio de Integração Acadêmica
"Das Montanhas de Minas ao Oceano: Os Caminhos da Ciência para um Futuro Sustentável"
SIA UFV 2025

Universidade Federal de Viçosa

Machine Learning Eficiente com GPU e FPGA

Pedro Henrique Coura Pereira, Ricardo dos Santos Ferreira, Olavo Alves Barros Silva, Isabela de Castro Freitas

ODS9: Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação

Categoria: Pesquisa

Introdução

Esse trabalho consiste na avaliação e desenvolvimento de técnicas de melhora de desempenho na fase de treinamento, por meio da paralelização utilizando GPU, e na fase de inferência, por meio de quantização e poda utilizando o TreeLUT no contexto de programação de circuitos reprogramáveis conhecidos como FPGAs (Field-Programmable Gate Arrays), de algoritmos de machine learning.

Objetivos

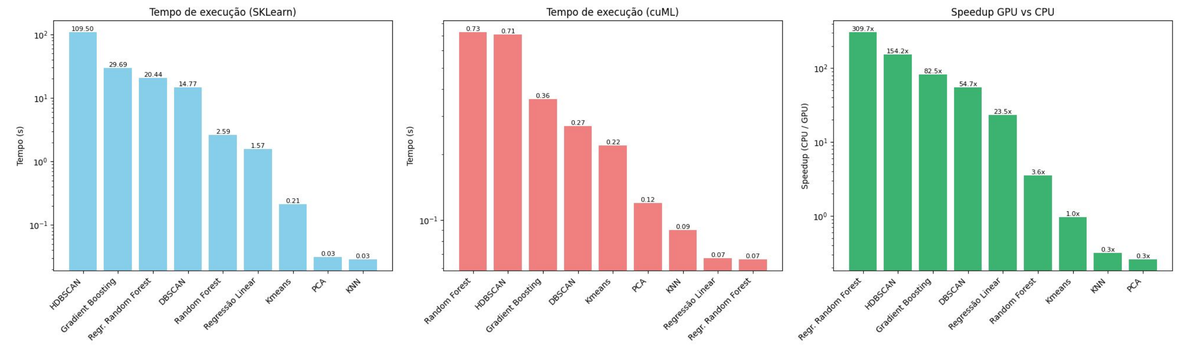
O objetivo deste trabalho é investigar estratégias para tornar o aprendizado de máquina mais eficiente, focando na aceleração do processamento e na otimização da inferência. Busca-se analisar como a execução em GPU pode reduzir o tempo de processamento de grandes volumes de dados e, simultaneamente, como a conversão de árvores de decisão em TreeLUT para FPGA, aliada a técnicas de poda, pode simplificar modelos para reduzir seu custo de inferência.

Material e Métodos ou Metodologia

Foram utilizados datasets reais e sintéticos para avaliar algoritmos de aprendizado de máquina em GPU com a biblioteca cuML/CUDA, medindo acurácia e tempo de execução. Paralelamente, árvores de decisão foram convertidas em TreeLUT para FPGA, aplicando poda para reduzir profundidade e gerar código C++ otimizado para inferência em hardware reconfigurável.

Resultados e/ou Ações Desenvolvidas

A paralelização garantiu redução significativa no tempo de treinamento na dos algoritmos testados, com as exceções sendo Kmeans, PCA e KNN, que aparentam ser mais difíceis de paralelizar por se basearem em operações iterativas e/ou que acessam frequentemente a memória, o que limita o aproveitamento da GPU.
A aplicação do TreeLUT foi capaz de quantizar os modelos sem perda significativa na acurácia, o que demonstra viabilidade do desenvolvimento de sistemas FPGAs eficientes e precisos.
Este trabalho também consistiu em desenvolver scripts em Python que convertem árvores no formato utilizado no TreeLUT para código CUDA C++ equivalente, de forma a facilitar implementações futuras.



Conclusões

A paralelização por meio de GPU é uma estratégia que pode acelerar significativamente o tempo de treinamento de algoritmos de machine learning, entretanto, sua eficácia varia de acordo com o algoritmo, e em alguns casos pode não trazer ganhos. Para acelerar a fase de inferência, FPGAs podem ser muito úteis e é possível quantizar um modelo por meio do TreeLUT sem perda significativa de acurácia.

Bibliografia

KHATAEI, A.; BAZARGAN, K. TreeLUT: An efficient alternative to deep neural networks for inference acceleration using gradient boosted decision trees. arXiv preprint arXiv:2501.01511, 2025.

Apoio Financeiro

CAPES FAPEMIG CNPq