TÉCNICAS DE OTIMIZAÇÃO DE REDES NEURAIS PROFUNDAS PARA APLICAÇÕES EM SISTEMAS EMBARCADOS

William H. A. Martins [1], **Rafael B. C. Lima** [2]

RESUMO

O Aprendizado de Máquina (Machine Learning - ML) e, em especial, o Aprendizado Profundo (Deep Learning - DL), têm se destacado como tecnologias cruciais na Inteligência Artificial (AI) devido à sua capacidade de melhorar continuamente a partir de grandes quantidades de dados. Embora a execução desses modelos geralmente exija recursos computacionais significativos, como GPUs e TPUs, a crescente demanda por aplicações com baixa latência e alta privacidade tem impulsionado o desenvolvimento de técnicas para executar modelos de ML em dispositivos embarcados. Este trabalho investiga métodos de compressão, quantização e otimização de modelos de redes neurais para torná-los mais eficientes e viáveis em dispositivos com recursos limitados, como microcontroladores de ultrabaixa potência. Ao explorar essas técnicas, o estudo visa aprimorar a execução de modelos de ML em dispositivos embarcados, contribuindo para a expansão do paradigma TinyML e o processamento de dados de dispositivos IoT, promovendo benefícios como baixa latência, eficiência energética e maior segurança dos dados.

Palavras-chave: Quantização, Deep Learning, TinyML, Dispositivos Embarcados, Otimização de Modelos.

[1] Aluno de Engenharia Elétrica, Departamento de Engenharia Elétrica e Informática, UFCG, Campina Grande, PB, e-mail: william.martins@ee.ufcg.edu.br

[2] D.Sc., Professor Adjunto, Departamento de Engenharia Elétrica e Informática, UFCG, Campina Grande, PB, e-mail: rafael.lima@dee.ufcg.edu.br

TÉCNICAS DE OTIMIZAÇÃO DE REDES NEURAIS PROFUNDAS PARA APLICAÇÕES EM SISTEMAS EMBARCADOS

ABSTRACT

Machine Learning (ML), and particularly Deep Learning (DL), have emerged as crucial technologies in Artificial Intelligence (AI) due to their ability to continuously improve from large amounts of data. While the execution of these models typically requires significant computational resources, such as GPUs and TPUs, the growing demand for applications with low latency and high privacy has driven the development of techniques to run ML models on embedded devices. This work investigates methods of compression, quantization, and optimization of neural network models to make them more efficient and viable on resource-constrained devices, such as ultra-low-power microcontrollers. By exploring these techniques, the study aims to enhance the execution of ML models on embedded devices, contributing to the expansion of the TinyML paradigm and the processing of data from IoT devices, promoting benefits such as low latency, energy efficiency, and greater data security.

Keywords: Quantization, Deep Learning, TinyML, Embedded Devices, Model Optimization.