Métodos Avançados de TinyMLOPs: Implementação de Aprendizado de Máquina em Sistemas Embarcados

Thommas Flores, Daniel Costa e Ivanovitch Silva

28 de julho de 2025

1 Justificativa

O minicurso "Métodos Avançados de TinyMLOPs: Implementação de Aprendizado de Máquina em Sistemas Embarcados"surge como uma resposta essencial à crescente demanda por capacitação em técnicas de machine learning (aprendizado de máquina) para dispositivos embarcados, impulsionada pela expansão da edge computing (computação de borda).

Este minicurso visa capacitar profissionais e pesquisadores a superar as barreiras técnicas da implementação de machine learning em dispositivos de baixa potência, como o ESP32, por meio de uma abordagem teórico-prática. Ao final, os participantes estarão preparados para aplicar estratégias modernas de quantização e deployment (implantação), contribuindo para o avanço de aplicações em edge AI (IA de borda), automação industrial e dispositivos inteligentes. Este é um passo crucial para impulsionar a inovação em soluções computacionais eficientes e autônomas, alinhadas às demandas do cenário tecnológico atual.

Com foco na quantização inteligente — incluindo Quantization-Aware Training (QAT, ou Treinamento com Consciência de Quantização) e Post-Training Quantization (PTQ, ou Quantização Pós-Treinamento) — e na compressão evolutiva de modelos de aprendizado de máquina, este minicurso está centrado na utilização do framework TensorFlores. O TensorFlores é um framework recentemente desenvolvido pelo grupo de pesquisa Conecta2AI da Universidade Federal do Rio Grande do Norte (UFRN) com colaboração do Laboratory of Emerging Smart Systems (LES2) da Faculdade de Engenharia da Universidade do Porto (FEUP), que permite a otimização de redes neurais Multilayer Perceptron (MLP, ou Perceptron Multicamadas) sem perda significativa de precisão. Além disso, o framework automatiza a conversão para código C++, garantindo compatibilidade com uma ampla gama de plataformas, desde microcontroladores até sistemas mais robustos. Essa combinação de técnicas avançadas e geração automatizada de código posiciona o TensorFlores como uma ferramenta indispensável para o desenvolvimento de IA embarcada eficiente e acessível.

1.1 Objetivo Geral

O objetivo geral do minicurso é capacitar os participantes no desenvolvimento e implementação de modelos de machine learning otimizados para dispositivos embarcados. Para isso, serão abordados os fundamentos da quantização, tanto pós-treinamento quanto durante o treinamento, além do processo de conversão de modelos para execução eficiente em microcontroladores, utilizando a plataforma TensorFlores.

1.2 Objetivos Específicos

- Explorar os princípios matemáticos subjacentes à quantização de modelos de *machine learning*, proporcionando aos participantes uma compreensão aprofundada dos métodos de redução de precisão e sua aplicação em dispositivos embarcados.
- Capacitar os participantes a aplicar técnicas avançadas de compressão de modelos de *machine learning*, com enfoque no método de quantização pós-treinamento e treinamento consciente de quantização, para otimizar o desempenho e a eficiência energética em ambientes de computação de recursos limitados.

- Fornecer oportunidades práticas para os participantes implementarem os conceitos aprendidos em laboratórios práticos, permitindo-lhes desenvolver e testar modelos de TinyML otimizados para dispositivos *edge*, consolidando assim o conhecimento teórico em habilidades aplicáveis.

2 Público-alvo, Pré-requisitos e Quantidade de Vagas

Este minicurso proporcionará uma imersão completa no universo do *Machine Learning* e dos Sistemas Embarcados, fornecendo aos participantes uma base sólida para explorar essas áreas de forma prática e inovadora. Este minicurso conta com uma oferta de **20 vagas**.

2.1 Público-Alvo

Estudantes, professores e pesquisadores dos cursos de Engenharia e Computação ou áreas afins, assim como todos aqueles que possuam interesse na temática abordada.

2.2 Pré-Requisitos

É imprescindível possuir conhecimento mínimo em programação, especialmente em Python e C++ (na plataforma IDE Arduino).

3 Ementa do curso

O curso contará com uma carga horária de 3 horas,

3.1 Parte Teórica (1h)

3.1.1 Introdução aos Conceitos de TinyML (20 minutos)

- Definição de TinyML e sua importância em dispositivos embarcados;
- Aplicações práticas de TinyML na indústria e na vida cotidiana;
- Desafios e oportunidades ao implementar modelos de machine learning em dispositivos com recursos limitados.

3.1.2 Métodos Avançados de Quantização Pós Treinamento (20 minutos)

- Visão geral dos métodos de quantização pós treinamento;
- Exemplos de casos de uso e melhores práticas em TinyML.

3.1.3 Métodos Avançados de Treinamento Consciente de Quantização (20 minutos)

- Visão geral dos métodos de quantização durante o treinamento;
- Exemplos de casos de uso e melhores práticas em TinyML.

3.2 Parte Prática (2h)

3.2.1 Configuração do Ambiente de Desenvolvimento para TensorFlores (50 min)

- Introdução ao framework TensorFlores e suas funcionalidades;
- Fluxo de trabalho no TensorFlores: Treinamento, quantização e conversão para código embarcado;
- Geração de código C++ e integração com microcontroladores;
- Instalação e configuração do TensorFlores no ambiente de desenvolvimento;
- Configuração de ferramentas e recursos necessários para o desenvolvimento de modelos de TinyML.

3.2.2 Configuração do Ambiente de Desenvolvimento (50 min hora)

- Guia prático para treinar, quantizar e embarcar modelos de aprendizado de máquina *Multilayer perceptron* pós-treinamento;
- Guia prático para treinar, quantizar e embarcar modelos de aprendizado de máquina *Multilayer perceptron* durante o treinamento.

3.2.3 Desafios e Discussões sobre Implementação de TinyML em Projetos Reais (20 minutos)

- Identificação e discussão dos desafios comuns enfrentados ao implementar TinyML em projetos reais;
- Compartilhamento de experiências e dicas práticas para superar esses desafios;
- Oportunidade para os participantes compartilharem seus próprios projetos e desafios.

Espera-se que ao final do minicurso, os participantes estejam equipados com o conhecimento teórico e prático necessário para desenvolver e implementar modelos de *machine learning* otimizados para dispositivos embarcados usando TensorFlores.

4 Minicurrículo dos Autores

4.1 Thommas Flores

Graduado em Engenharia Elétrica pela Universidade Federal da Paraíba em 2018, com Mestrado em Engenharia Elétrica pela mesma universidade em 2021, com foco em controle adaptativo. Atualmente, é doutorando em Engenharia Elétrica e de Computação na Universidade Federal do Rio Grande do Norte. Sua pesquisa foca em IA embarcada, MLOps, Internet dos Veículos Inteligentes e Controladores Inteligentes para aplicações industriais.

4.2 Daniel Costa (Membro Sênior, IEEE)

Professor Assistente no Departamento de Engenharia Elétrica e de Computação (DEEC) da Faculdade de Engenharia (FEUP) da Universidade do Porto. Autor ou coautor de mais de 100 artigos científicos em revistas e conferências internacionais, com temas em Cidades Inteligentes, Internet das Coisas, Redes de Sensores, Sistemas Embarcados e Sensoriamento Multimídia.

4.3 Ivanovitch Silva (Membro Sênior, IEEE)

Professor Associado no Departamento de Engenharia de Computação e Automação (DCA) da UFRN e, em 2021, foi Professor Visitante na Università degli Studi di Brescia, Itália. Graduado em Engenharia Elétrica e de Computação pela Universidade Federal do Rio Grande do Norte (UFRN) em 2006, com Mestrado (2008) e Doutorado (2013) pela mesma instituição, com co-participação da Universidade do Porto, Portugal. Seus interesses de pesquisa são em IA embarcada, MLOps, Internet dos Veículos Inteligentes e Cidades Sustentáveis.

5 Material e Recursos

Para ministrar o minicurso de Machine Learning e Sistemas Embarcados, é importante contar com uma série de recursos e materiais para garantir uma experiência eficaz e enriquecedora para os participantes. Abaixo, listo alguns dos recursos e materiais essenciais:

- Computadores para os participantes, preferencialmente com acesso à internet e capacidade para executar softwares de desenvolvimento, como Python IDEs e Arduino IDE
- Projetor multimídia ou televisão para a exibição de *Slides* ou apresentações sobre os conceitos teóricos e práticos de Machine Learning e Sistemas Embarcados.