Tiny-MLOps: a framework for orchestrating ML applications at the far edge of IoT systems

Machine Learning models are usually hosted on Linux-based single-board computers. However, these single-board computers are expensive and consume much power, limiting their use in many cases.
32-bit microcontrollers are cheaper and require little power to operate, making them a cost-effective and energy-efficient alternative. However, these microcontrollers have limited resources and run on non-Linux operating systems. Managing ML tasks on these resource-limited devices is challenging with current tools.
This paper introduces the Tiny-MLOps framework, which adapts standard ML management practices to work with far-edge devices(32-bit microcontrollers). The framework adjusts each phase of the typical ML process to fit the limited resources of IoT devices.
The framework was tested on IoT sensors mounted on an industrial machine to detect problems. They showed how to update ML-based problem detection models on far-edge devices.

Strengths:

- It uses low-cost, low-power microcontrollers, making it affordable and energy-saving for IoT applications.
- This framework adjusts each phase of the MLOps loop to fit the limited resources of IoT devices, allowing them to be automatically re-configured and adapted during operation.

Weaknesses:

- The deployment of a new model may take hours to complete because the model is stored in a non-optimized format for the target device due to its limitations.
- The system may be sensitive to false positives since the pool of models may fail in the detection of anomalies.

- Far edge devices have limited power, memory, and storage, limiting the complexity of ML models they can run.

Open Issues:

- There is an issue related to optimizing the deployment of models.
- How to select the best ML model for deployment on resource-constrained devices.
- One of the issues which is not mentioned is integrating the Tiny-MLOps framework with existing IoT infrastructures and workflows.

The paper's relevance to our project goals:

- The paper addresses the challenges of deploying ML models on resource-constrained far edge devices (similar to mobile phones)
- The authors provide a foundation for optimizing resource consumption.