Scalable Reasoning Framework for Edge Devices

Master's Thesis Proposal - Fall, Winter and Spring 2025/26

Author: Alex Miller, Stanford University / Cal Poly
Advisor: Franz J. Kurfess

Abstract

This thesis investigates a scalable reasoning framework designed for AI deployment on resource-constrained edge devices. The research focuses on optimizing inference pipelines and knowledge representation under memory and latency constraints. The proposed system aims to enable efficient, real-time reasoning on devices such as drones, IoT sensors, and mobile robots.

Thesis Overview

AI reasoning typically depends on high-compute environments, limiting real-world deployment. This project proposes a lightweight framework that brings logical reasoning to edge devices. It explores compression techniques and modular architectures to enable symbolic and probabilistic reasoning within restricted computational budgets.

Background

Edge computing offers benefits in latency, privacy, and autonomy but poses challenges for running AI models locally. Existing frameworks like TensorFlow Lite optimize deep networks but not reasoning layers. This work extends compact model design into logical reasoning contexts.

Related Work

Research on TinyML and edge AI (Banbury et al., 2021) has addressed model size and efficiency but not symbolic reasoning. The proposed system draws on neuromorphic computing principles and rule-based compression to achieve lightweight inference.

Contributions

1. A compact reasoning engine for edge deployment.
2. Methods for adaptive knowledge compression.
3. Experimental validation on real-world edge devices.

Thesis Question / Hypothesis

Hypothesis: Modular, compressed reasoning architectures can achieve near-cloud-level inference accuracy on edge devices with less than 20% of computational cost.

Research Goal and Methodology

The project involves designing an optimized reasoning engine combining symbolic logic and probabilistic inference. Experiments will be conducted on Raspberry Pi and NVIDIA Jetson boards. Evaluation will compare latency, accuracy, and energy consumption against centralized systems.

Evaluation and Validation Criteria

Benchmarks include reasoning time, energy efficiency, and scalability with growing rule sets. Validation will include ablation studies on model compression and caching mechanisms.

Expected Outcomes and Significance

The expected outcome is a generalizable reasoning framework that democratizes AI capabilities across distributed devices. The research could impact autonomous robotics, IoT, and real-time analytics by enabling local decision-making with reduced dependency on cloud infrastructures.