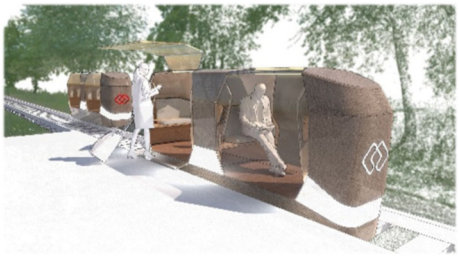
Comprehensive Analysis of Model Deployment and Optimization Strategies for Edge Devices

Future transport will focus on sustainability and efficiency, with rail as a key pillar. The MONOCAB, a gyro-stabilized monorail, aims to revive rural rail lines through full automation enabled by advanced sensors and machine learning. This innovation ensures safe, autonomous operation and supports a demand-responsive, integrated, and sustainable transport network, reinforcing rail's role in future mobility.



Project Aim

As part of the enableATO project, multiple AI algorithms will be developed for signal and sign recognition, semantic scene understanding, object detection, and intention recognition to enable fully autonomous operation of the MONOCAB. Running all models simultaneously onboard can create computational constraints, making deployment and optimization strategies essential to ensure real-time performance and system reliability. This project focuses on efficient model deployment and optimization on edge devices, allowing the MONOCAB to safely and autonomously navigate rural rail lines, supporting sustainable, demand-responsive, and integrated transport solutions.

The objectives of this project are as follows:

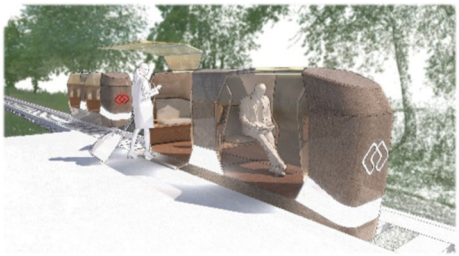
- **Research Deployment and Optimization Strategies:** Investigate state-of-the-art techniques for deploying AI models on edge devices, including quantization, pruning, and hardware-specific optimizations.
- **Model Selection and Performance Optimization:** Select two to three suitable pre-trained models for tasks such as signal recognition, object detection, and semantic segmentation, and apply optimization strategies to enhance their efficiency and performance on the target platform.
- **Comparative Analysis:** Evaluate and compare the optimized models in terms of accuracy, inference speed, memory usage, and robustness, identifying the most effective deployment approach for real-time autonomous operation.

Supervisor

M.Sc. Omar Gamal, omar.gamal@th-owl.de

Prof. Ulrich Büker, ulrich.bueker@th-owl.de

Umfassende Analyse von Modellbereitstellungs- und Optimierungsstrategien für Edge-Geräte



Der zukünftige Verkehr wird sich auf Nachhaltigkeit und Effizienz konzentrieren, wobei die Schiene eine zentrale Säule bildet. Der MONOCAB, eine gyrostabilisierte Einschienenbahn, hat das Ziel, ländliche Bahnstrecken durch vollständige Automatisierung mithilfe fortschrittlicher Sensorik und maschinellen Lernens wiederzubeleben. Diese Innovation gewährleistet einen sicheren, autonomen Betrieb und unterstützt ein nachfrageorientiertes, integriertes und nachhaltiges Verkehrsnetz, das die Rolle der Schiene in der Mobilität der Zukunft stärkt.

Projektziel

Im Rahmen des enableATO-Projekts werden mehrere KI-Algorithmen für Signal- und Schildererkennung, semantisches Szenenverständnis, Objekterkennung und Intentionserkennung entwickelt, um den vollautonomen Betrieb des MONOCAB zu ermöglichen. Da das gleichzeitige Ausführen aller Modelle an Bord rechnerische Einschränkungen verursachen kann, sind Einsatz- und Optimierungsstrategien entscheidend, um Echtzeitfähigkeit und Systemzuverlässigkeit sicherzustellen. Das Projekt konzentriert sich auf die effiziente Modellbereitstellung und -optimierung auf Edge-Geräten, wodurch das MONOCAB sicher und autonom auf ländlichen Bahnstrecken navigieren kann und somit nachhaltige, nachfrageorientierte und integrierte Verkehrslösungen unterstützt.

Die Ziele dieses Projekts sind wie folgt:

- **Forschung zu Bereitstellungs- und Optimierungsstrategien:** Untersuchung modernster Verfahren zur Bereitstellung von KI-Modellen auf Edge-Geräten, einschließlich Quantisierung, Pruning und hardware-spezifischer Optimierungen.
- **Modellauswahl und Leistungsoptimierung:** Auswahl von zwei bis drei geeigneten vortrainierten Modellen für Aufgaben wie Signalerkennung, Objekterkennung und semantische Segmentierung, sowie Anwendung von Optimierungsstrategien, um deren Effizienz und Leistung auf der Zielplattform zu verbessern.
- **Vergleichende Analyse:** Bewertung und Vergleich der optimierten Modelle hinsichtlich Genauigkeit, Inferenzgeschwindigkeit, Speichernutzung und Robustheit, um den effektivsten Bereitstellungsansatz für den echtzeitfähigen autonomen Betrieb zu identifizieren.

Betreuer

M.Sc. Omar Gamal, omar.gamal@th-owl.de

Prof. Ulrich Büker, ulrich.bueker@th-owl.de