EdgeGuard: Real-Time Industrial Safety Monitoring using NPU Acceleration on NXP i.MX 8M Plus

Problem Definition

In high-risk industrial environments (construction sites, factories, chemical plants), compliance with Personal Protective Equipment (PPE) regulations is crucial for preventing injuries. Traditional monitoring relies on human supervision, which is intermittent and prone to error, or cloud-based analytics, which suffer from latency and bandwidth issues. A delayed alert regarding a missing safety helmet can result in a fatal accident.

Project Scope and Solution

We propose **EdgeGuard**, an autonomous vision system deployed on the NXP i.MX 8M Plus development board. The system will process video feeds locally to detect safety violations in real-time.

- **Objective:** Detect workers and verify the presence of safety helmets (Hard Hats).
- **Technology:** The solution leverages the integrated **VeriSilicon Neural Processing Unit (NPU)** to execute quantized Deep Learning models (TensorFlow Lite) at high speed, ensuring immediate feedback without internet dependency.

Challenges in Traditional Practices

- **Latency:** Sending video to the cloud introduces delays (hundreds of milliseconds to seconds), unacceptable for immediate safety interlocks.
- **Bandwidth Cost:** Streaming 24/7 HD video from multiple cameras is expensive and strains the network.
- **Privacy:** Streaming employee footage to external servers raises GDPR and privacy concerns. Our Edge AI solution ensures video data never leaves the device.

Feasibility and Preliminary Results

We have conducted initial tests on the NXP i.MX 8M Plus EVK using the internal NPU (`libvx_delegate`).

- **Benchmarking:** Running a MobileNet V1 (int8 quantized) model resulted in an average inference time of **2.96 ms** per frame.
- **Throughput:** This performance indicates a theoretical throughput of over 300 FPS, providing ample computational headroom to implement complex object detection logic (SSD - Single Shot Detector) while maintaining real-time performance (30 FPS).

Team Roles

- **Bucătaru Alexandra:** Data Engineering. Responsible for dataset curation (collecting images of PPE), data augmentation, and training the custom object detection model using Transfer Learning.
- **Răşcanu Robert:** Edge Deployment Architect. Responsible for hardware setup, NPU driver integration, optimizing the inference pipeline (Python/C++), and developing the alert logic on the i.MX 8M Plus.

Literature Review

1. The Shift to Edge Intelligence The paradigm of AI is shifting from the Cloud to the Edge. As highlighted by Zhou et al. in *"Edge Intelligence: Paving the Last Mile of Artificial Intelligence"* [2], the proximity of computation to the data source significantly reduces latency and energy consumption. For industrial safety, where reaction time is critical, this shift is mandatory.

2. Hardware Acceleration in Embedded Systems Running Deep Learning models on standard CPUs (like the Cortex-A53) is often insufficient for video processing. The *ARM Edge AI Technology Report (2023)* emphasizes the role of heterogeneous computing. Our preliminary tests confirm this: while CPU inference often exceeds 50-100ms for detection tasks, utilizing the dedicated NPU on the i.MX 8M Plus reduces inference time to under 3ms for classification tasks. This aligns with the findings of Wang et al. [1], who argue that specialized hardware accelerators are the key enablers for "Real-Time Edge Intelligence."

3. Model Optimization Techniques To fit complex Convolutional Neural Networks (CNNs) onto edge devices, optimization is required. MIT News [4] reports that techniques like **Post-Training Quantization (PTQ)**—converting 32-bit floating-point weights to 8-bit integers—can reduce model size by 4x with negligible accuracy loss. We plan to utilize TensorFlow Lite's quantization toolkit to adapt our PPE detection models for the VeriSilicon NPU, ensuring robust performance in resource-constrained environments.

References

[1] X. Wang et al., "Convergence of Edge Computing and Deep Learning," IEEE Comm. Surveys & Tutorials, 2020.

[2] Z. Zhou et al., "Edge Intelligence: Paving the Last Mile of AI," IEEE, 2019.

[3] "Training Machine Learning models at the Edge: A Survey," arXiv:2403.02619v3.