Trans-Precision Neural Network Deployment for Low-Power Embedded Systems

The methodology efficiently deploys and accelerates floating-point neural networks on embedded systems, optimizing performance, energy consumption, and hardware utilization.

Abstraction levels:

1. Model deployment

2. **System infrastructure**

3. Streamlined acceleration

4. Optimized processing

HW/SW co-design framework

Application

Machine learning library

TensorFlow Lite core

TensorFlow Lite delegate

Hardware abstraction layer

Hardware drivers

Board support package

Hardware platform (Zynq SoC)

Programmable Logic (PL)	Processing System (PS)
Tensor Processor	ARM Cortex-A9