Machine Learning

For Embedded Devices

Presented by:
Manuel Capel

Fablab Winti
10. November 2023

Table of Content

- ➔ What is Machine Learning
- ➔ Unleashing the potential of embedded IoT devices
- ➔ Example of use-cases
- ➔ Model Optimization
- ➔ Model Workflow
- ➔ Introduction to EdgeImpulse

  Demo

What is Machine Learning

AI: Program that can learn how to perform a task requiring typical “human intelligence”

ML: AI algorithm learning to perform a precise task given training data.
Tasks:

- Classification: objects on images, disease…
- Regression: predicting stock exchange, conversion rate on a website…
- Reinforcement learning: games like chess or go, robots in an environment…
- Generative: text (ChatGPT), images (DALL-E)...

Deep Learning: ML algorithms using *neural networks*
Main Architectures:

- CNN (convolutional): especially for images
- RNN (recurrent): for sequence predictions (time series…)
- Transformers: for generative models

Unleashing Embedded Potential

IoT devices are everywhere:

- ~25 IoT devices in a home (source: Deloitte)
- >100 Micro-Controllers in a car, controlling engine, transmission etc.
- IoT: backbone of critical infrastructures (water, electricity distribution...)

Growing rapidly: 22bn devices in the world in 2022, 75bn expected in 2025 (source: singularity group)

Machine Learning on device vs. on cloud:

- ➢ Faster reaction (no latency)
- ➢ More secure (reduced attack surface)
- ➢ Not reliant on network connection
- ➢ More eco-friendly because less energy intensive

Contras:

- ➢ Deployment on the fleet (OTA: Over The Air)
- ➢ Limited computing capacities

Embedded ML Use Cases

Industrie 4.0
- Anomaly detection
- Predictive Maintenance
- Better energy efficiency
- QA (Quality Assurance) on production line

Health / wellbeing:
- Wearables: not only measuring, but also assessing and reacting
- Monitoring devices (cameras, movement / fatigue detection)

=> Better privacy due to minimized data transmission

Agriculture:
- Monitoring crop, stocks and watering, adapting temperature optimally
- Early detection of diseases
- Optimal use of resources

=> Better sustainability, speed, privacy...

Model Optimization

TinyML: Machine Leaning in the mW range
Challenges: cross-compilation, model compression

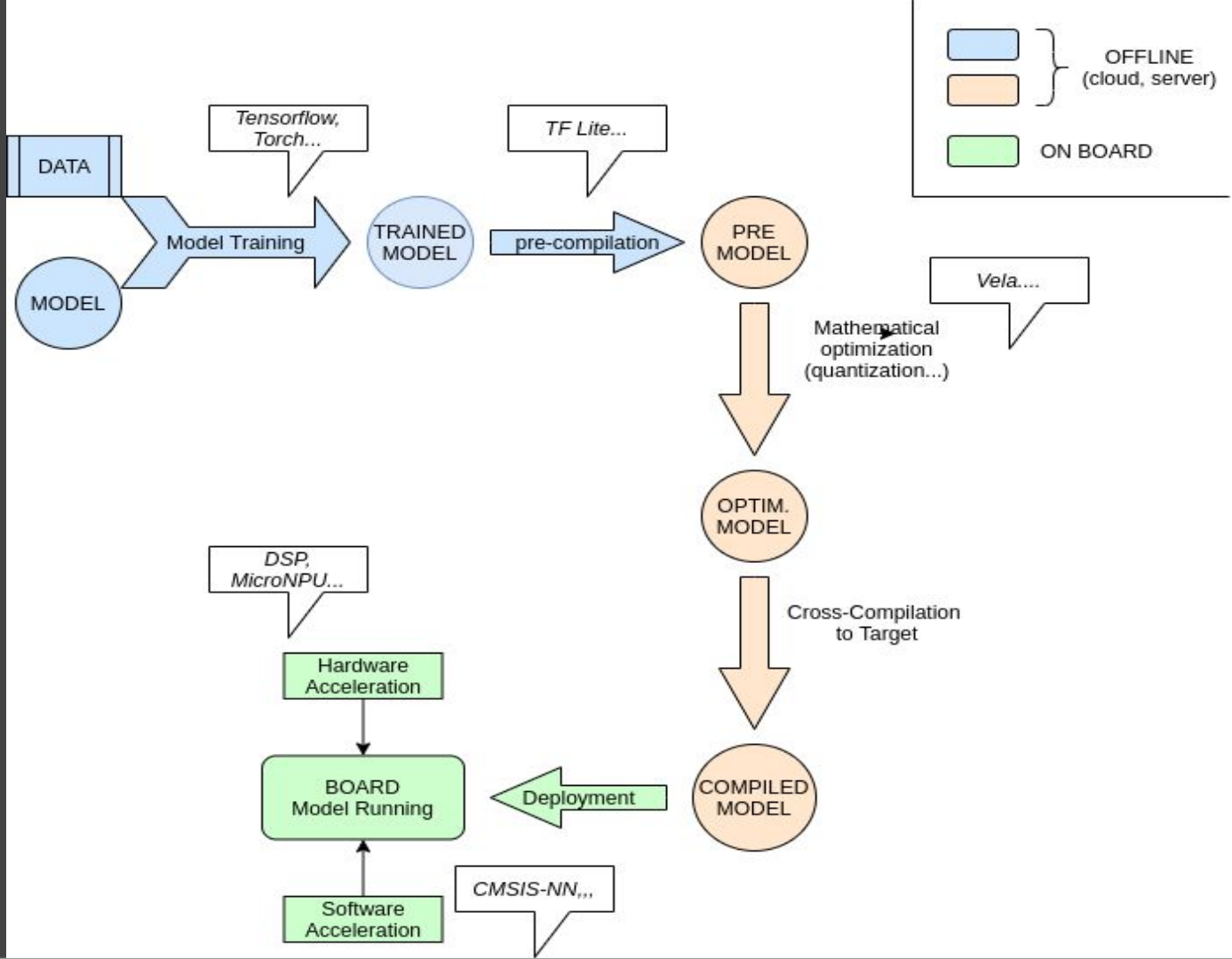
Mathematical methods:
- quantization: float => int: less memory and faster processing
- pruning: removing "dead" (zero) parameters
- memory alignment

Hardware acceleration: ARM has been leading:
- Helium (Cortex M series): Optimized vector computation
- DSP (Digital Signal Processing) modules
- MicroNPU (Neural Processing Unit): offload from CPU, compressed weights...

Software acceleration, for example:
- Vela (compiler, offline)
- CMSIS-NN (on-board)

Performance x25 - x1000 depending on task vs. "vanilla" tflite
// NB: different types of performance: energy consumption, time of inference...

Model Workflow



Demo

- ❑ Setting up device
- ❑ Connecting device
- ❑ Sampling Data
- ❑ Training Model
- ❑ Compile
- ❑ Deploy
- ❑ Enjoy!