ONNX on MCUs & TinyML devices

ROHIT SHARMA
rohit@ai-tech.systems

© confidential
Agenda

- ONNX, TinyML and MCUs
- TinyML on MCUs: Challenges and Opportunities
- TinyML: Frameworks, tools, and techniques
- TinyML Applications and Use Cases
  - Sensor ASL
  - Wake word detection
- deepSea: ONNX Compiler, Library & Framework
- Questions
What is ML in TinyML?

- Network
  - Cloud
  - Edge
- ML
  - Model
  - Training
  - Inference
- Hardware
  - IoT
  - MCU
- Benefits
  - Low Power
  - Low Latency
  - Security
What is TinyML?

Many Things including:

- Low Power ML apps
- Running on Embedded Systems
- With optional cloud connectivity
- Making Edge compute affordable
- With Privacy and Low latency
TinyML: Challenges and Opportunities

Explosion in Edge Markets & Verticals

Vision AI
NLP
Mobile

Bio Metric
Sensor AI
Industrial

Global market for CMOS image sensors 2013-2023

Source: IC Insights

World population
Annual Sensor Birth Rate

Internet bandwidth of the world 1990-2018
TinyML Hardware

Limited Resources

GPU (GPU Servers)
- $10^{12}$ FLOPS
- Multi boards GHz
- 64GB RAM
- 455 lb CO$_2$ emission

CPU (Mobile, Laptop)
- $10^9$ FLOPS
- Multi core GHz
- 4 GB RAM
- 45.4 lb CO$_2$ emission

MCU (Tiny Machines)
- $10^6$ FLOPS
- Single core MHZ
- 256K RAM
- 11.4k lb CO$_2$ emission
TinyML Apps

Increasing Power & Cost

Floating Point Ops per Sec

<10 mW

1W

10W

>100W

Micro Controllers

Wake word Detection

Image Recognition

Image Enhancement

Autonomous Vehicle

© confidential
Steps for TinyML app Development

- Gather Data
- Design Network
- Train Model
- Export Model
- Compile Model
- Design App
deepC: Compile ONNX Model


ONNX IR
- TF PB
- Keras h5
- PyTorch
- Caffe

Edge AoT Compiler
- ONNX Graph
- Compute Graph
- Schedule
- Optimize
- Codegen

Outputs
- Embedded C/C++
- Web Assembly
- Static Library
- Bare Metal Binary
- OS ELF
TinyML Tools: Library, Compiler, Framework and Platform

deepC: open source ONNX AoT Compiler

cAlnvas: tinyML platform

www.tinyML.studio

No Code
(Bring Your Own Model)

Low Code
(Bring Your Own Data)

Full Code
(Bring Your Own Resource)
End to End Platform
No Setup
Flexible Training
Cross Compilation
Pre Compiled Models
Free Tier
Use Case: aiSpeak for mute

Gesture to speech Technology
Use Case: aiSpeak for mute

Ref: https://cainvas.ai-tech.systems/use-cases/sign-language-sensor-app/

Accelerometer & Gyroscope Data → aiSpeak Neural Network → Top-5 word prediction → Next Word Predictor → Word to speech