The Era of TinyML on AIoT with Microcontrollers (MCUs)

Microcontrollers: low-cost, low-power, widely deployed

Wide Applications

Challenge: Memory Too Small to Hold DNNs

MCUNet: Tiny Deep Learning on IoT Devices

MCUNet: System-Algorithm Co-design

(a) Search NN model on an existing library, e.g., ProxylessNAS, MnasNet

(b) Tune deep learning library given a NN model, e.g., TVM

Efficient Neural Architecture

TinyNAS

MCUNet

TinyEngine

Efficient Compiler / Runtime

(c) MCUNet: system-algorithm co-design

1. TinyNAS: Two-Stage NAS for Tiny Memory

Memory/Storage Constraints

- Full Network Space
- Optimized Sub-space
- Model Specialization

- Analyzing FLOPs distribution of satisfying models in each search space:
  - Larger FLOPs -> Larger model capacity -> More likely to give higher accuracy

2. TinyEngine: Memory-Efficient Inference Library

TinyEngine allows for a fit a larger model at the same hardware resource by:
- Reduced peak memory usage
- Accelerated inference speed

- On-device deployment demo or visual wake word (person present or not)

Experimental Results

- System-algorithm co-design gives the best performance
  - ImageNet classification on STM32F469 MCU (320kB SRAM, 1MB Flash)
  - MicroTVM Tuned
  - MicroTVM Tuned
  - 1.6x faster

- MCUNet significantly outperforms existing solutions on visual/audio keyword spotting

- MCUNet automatically handles diverse hardware capacity by optimizing search spaces

Existing Methods Reduce Model Size, but not the Activation Size

- Existing network CANNOT fit the light memory constraints on MCU

ResNet-18

MobileNetV2-0.75

<table>
<thead>
<tr>
<th>Param (MB)</th>
<th>ResNet-18</th>
<th>MobileNetV2-0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Activation (MB)</td>
<td>1.8x</td>
<td>4.6x</td>
</tr>
</tbody>
</table>

- TinyEngine consistently improves on different networks

TinyEngine: A Memory-Efficient Inference Library

- Baseline: ARM CMSIS-NN
- MicroTVM Tuned

- Peak Mem (KB): TinyEngine vs. Baseline

- VWW Accuracy

- On-device deployment demo or visual wake word (person present or not)

- 75% Acc. on-device deployment demo or visual wake word (person present or not)

- 87% Acc. on-device deployment demo or visual wake word (person present or not)