MCUNetV1 & V2: On-Device Inference of Tiny Deep Learning on IoT Devices

Wei-Chen Wang
Song Han
Department of EECS
MIT
wweichen@mit.edu

Ji Lin, Wei-Ming Chen, Yujun Lin, Han Cai, John Cohn, Chuang Gan, Song Han
MIT, MIT-IBM Watson AI Lab

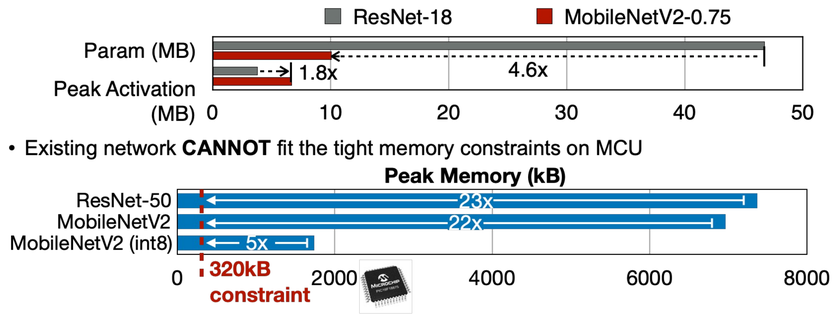
National Science Foundation

tinyml.mit.edu

Tiny Machine Learning (TinyML) Faces Challenge of Limited Memory

	ResNet Cloud AI	MobileNet Mobile AI		MCUNet Tiny AI
Memory	16GB	4GB	13,000x smaller →	320kB
Storage	~TB/PB	256GB		1MB

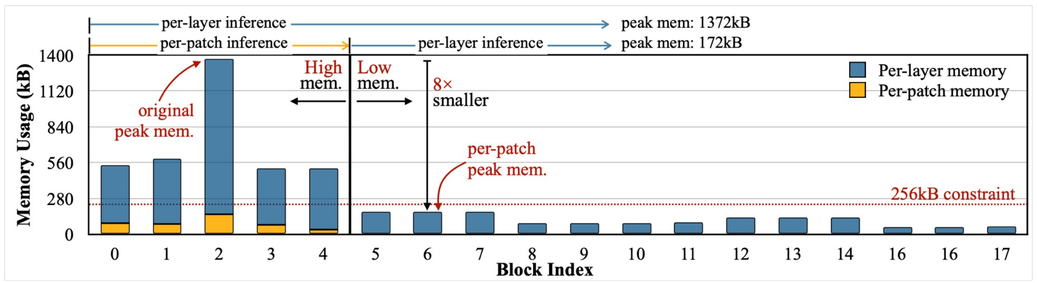
Challenge 1: Existing Methods Reduce Model Size, but not the Activation Size



- Existing network **CANNOT** fit the tight memory constraints on MCU

Challenge 2: Efficient CNNs Have Imbalanced Memory Distribution

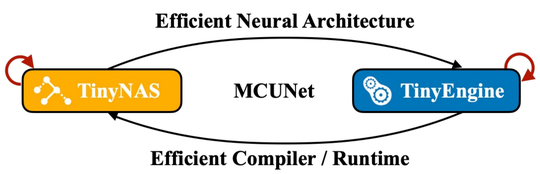
- Per-block memory usage of MobileNetV2
- Reduce memory usage of the **initial stages** -> Reduce the overall memory usage



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*

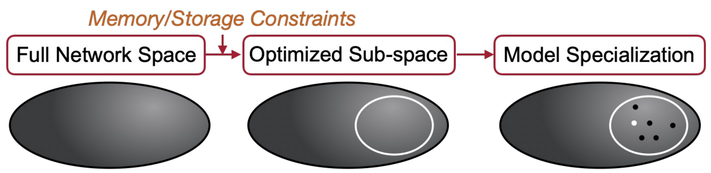


(c) *MCUNet*: system-algorithm co-design

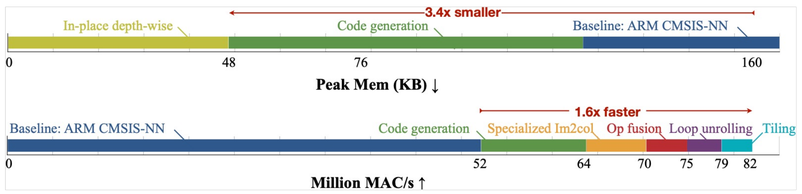
TinyNAS analyzes **FLOPs distribution** of satisfying models in each search space:
- Larger FLOPs -> Larger model capacity -> More likely higher accuracy

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

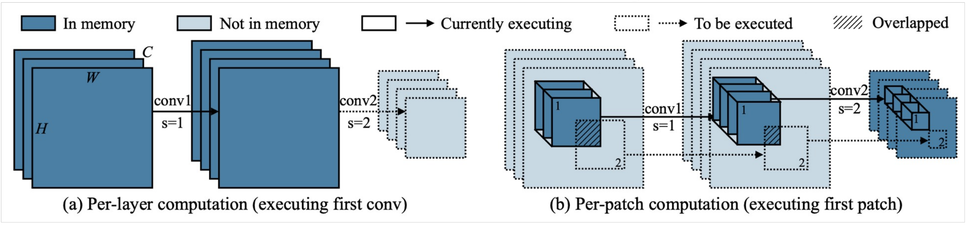
1. TinyNAS: Two-Stage NAS for Tiny Memory



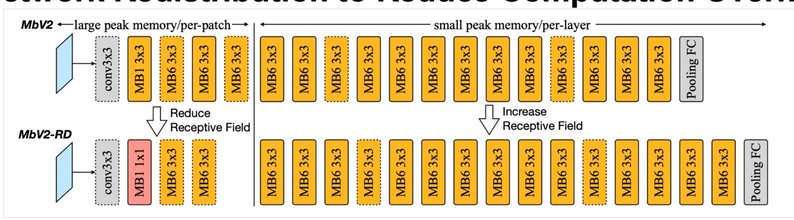
2. TinyEngine: Memory-Efficient Inference Library



3. Patch-based Inference to Break Memory Bottleneck

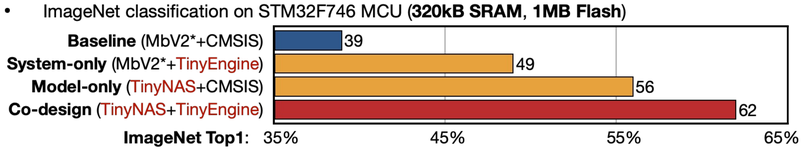


4. Network Redistribution to Reduce Computation Overhead



Experimental Results

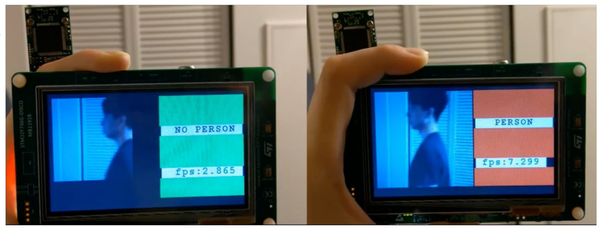
- **System-algorithm co-design** gives the best performance
  - ImageNet classification on STM32F746 MCU (**320kB SRAM**, **1MB Flash**)



- **On-device deployment** demo for visual wake word (person or no person)

MbV1+TF-Lite Micro

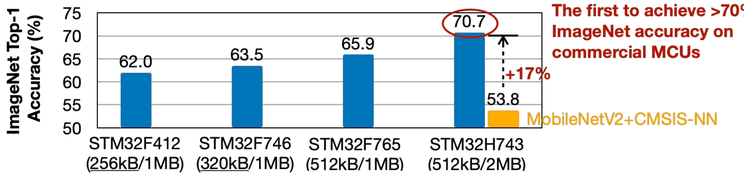
75% Acc.
2.9 FPS



MCUNet

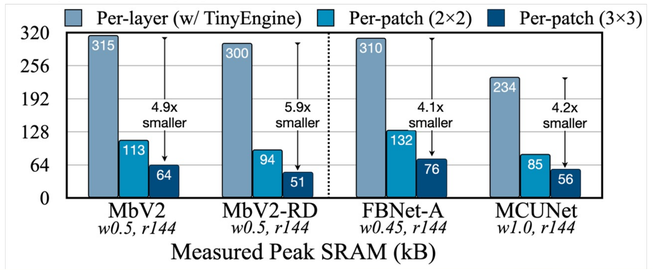
87% Acc.
7.3 FPS

- MCUNet automatically handles **diverse hardware capacity** by optimizing search spaces

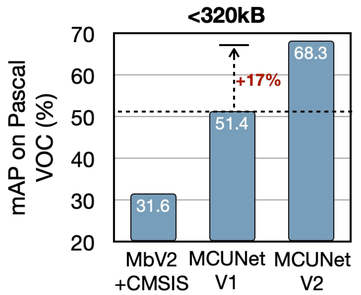


The first to achieve >70% ImageNet accuracy on commercial MCUs

- Reduce peak memory by 4-6x for off-the-shelf tiny models



- Face detection on WIDER Face
- More robust results at a smaller peak memory
- Patch-based inference allows for a **larger resolution**, improving detection performance



(a) RNNPool-Face-Quant (b) MCUNetV2

Acknowledgment to NSF Grant OAC-2117997