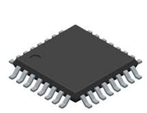
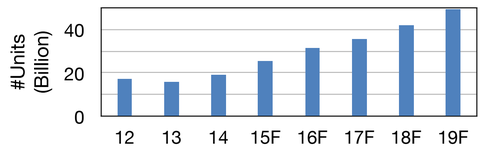
MCUNet: Tiny Deep Learning on IoT Devices

Ji Lin[1], Wei-Ming Chen[1,2], Yujun Lin[1], John Cohn[3], Chuang Gan[3], Song Han[1]
[1]MIT [2]National Taiwan University [3]MIT-IBM Watson AI Lab

The Era of TinyML on AIoT with Microcontrollers (MCUs)

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



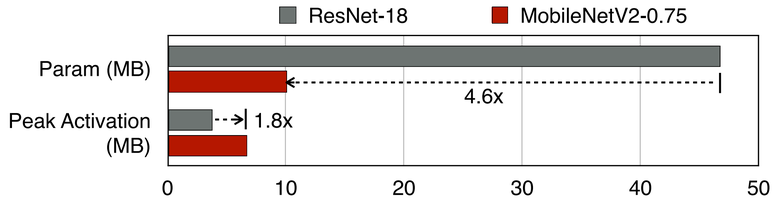


Wide Applications

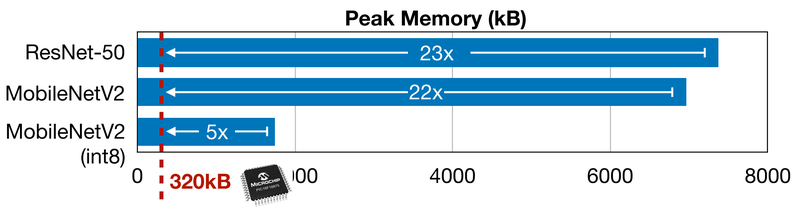
Challenge: Memory Too Small to Hold DNNs

	ResNet **Cloud AI**	MobileNet **Mobile AI**	MCUNet **Tiny AI**
Memory (Activation)	16GB	4GB ← 13,000x smaller	320kB
Storage (Weights)	~TB/PB	256GB	1MB

Existing Methods Reduce Model Size, but not the Activation Size



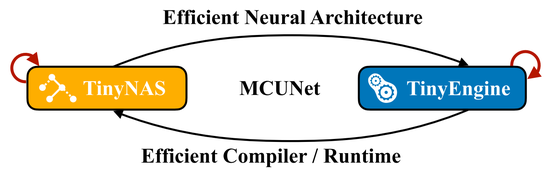
- Existing network **CANNOT** fit the tight memory constraints on MCU



MCUNet: System-Algorithm Co-design

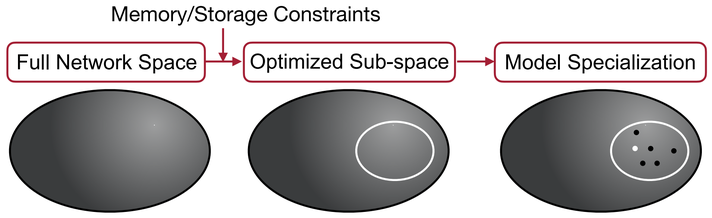
(a) Search NN model on an existing library. e.g., *ProxylessNAS, MnasNet*

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

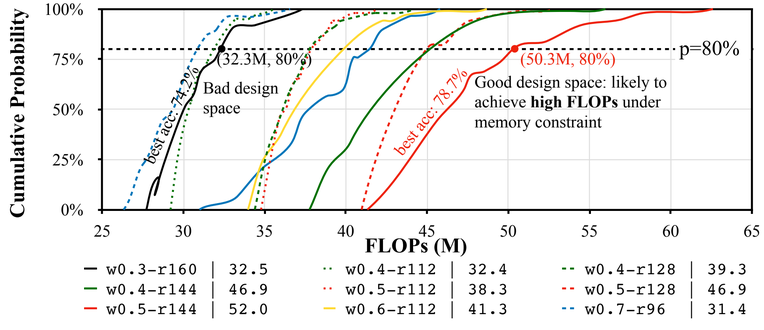


(c) *MCUNet*: system-algorithm co-design

1. TinyNAS: Two-Stage NAS for Tiny Memory



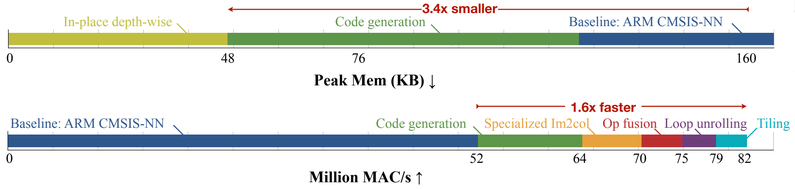
- Analyzing **FLOPs distribution** of satisfying models in each search space:
  Larger FLOPs -> Larger model capacity -> More likely to give higher accuracy



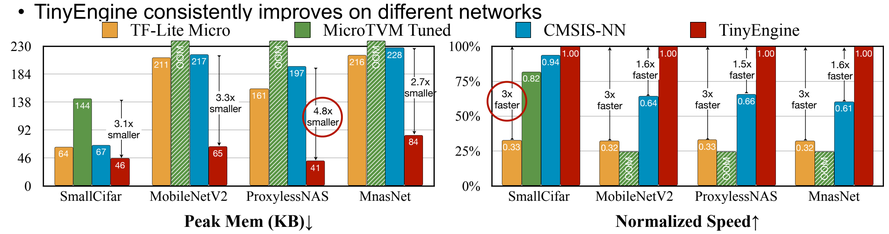
2. TinyEngine: Memory-Efficient Inference Library

TinyEngine allows to fit a larger model at the same hardware resource by:

- Reduced peak memory usage
- Accelerated inference speed

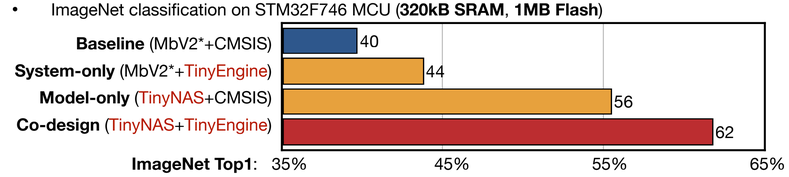


- TinyEngine consistently improves on different networks

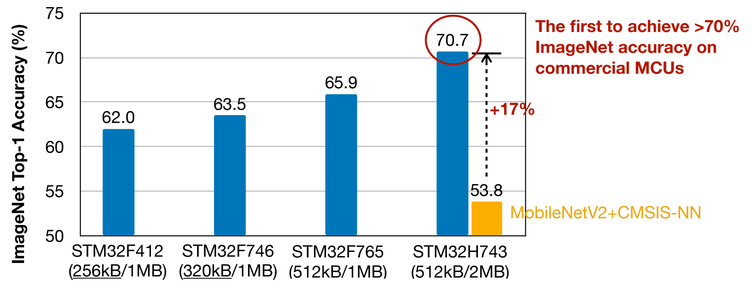


Experimental Results

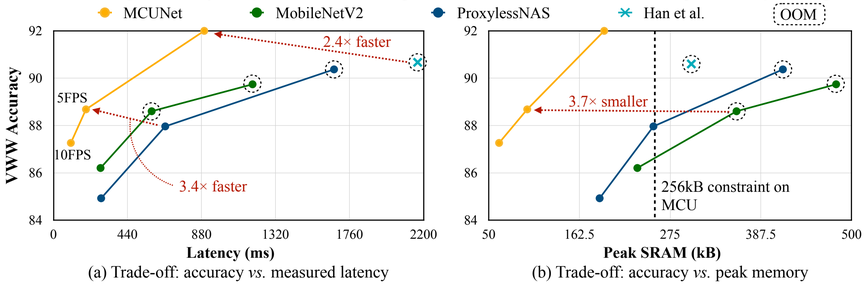
- **System-algorithm co-design** gives the best performance
  - ImageNet classification on STM32F746 MCU (**320kB SRAM, 1MB Flash**)



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



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



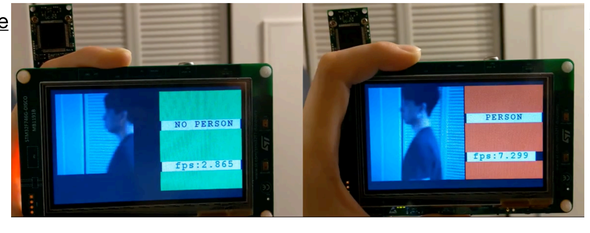
(a) Trade-off: accuracy *vs.* measured latency

(b) Trade-off: accuracy *vs.* peak memory

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

MbV1+TF-Lite Micro

75% Acc.
2.9FPS



MCUNet

87% Acc.
7.3 FPS