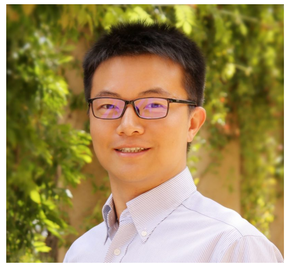
MIT HAN LAB

Hardware, AI, and Neural-nets
open source, co-design
http://github.com/mit-han-lab

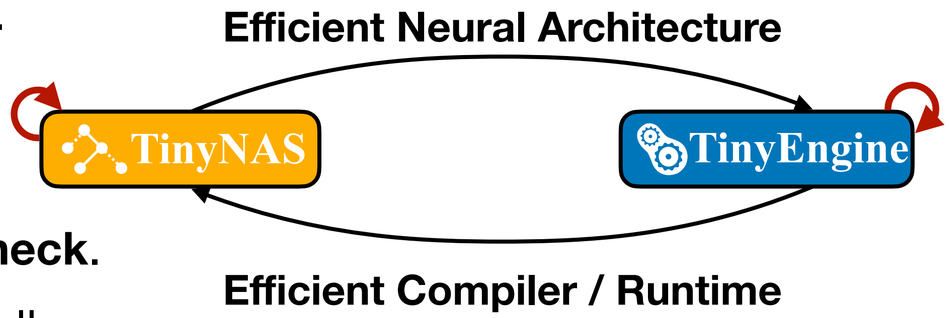
MCUNet: Model Compression and Tiny On-Device Learning

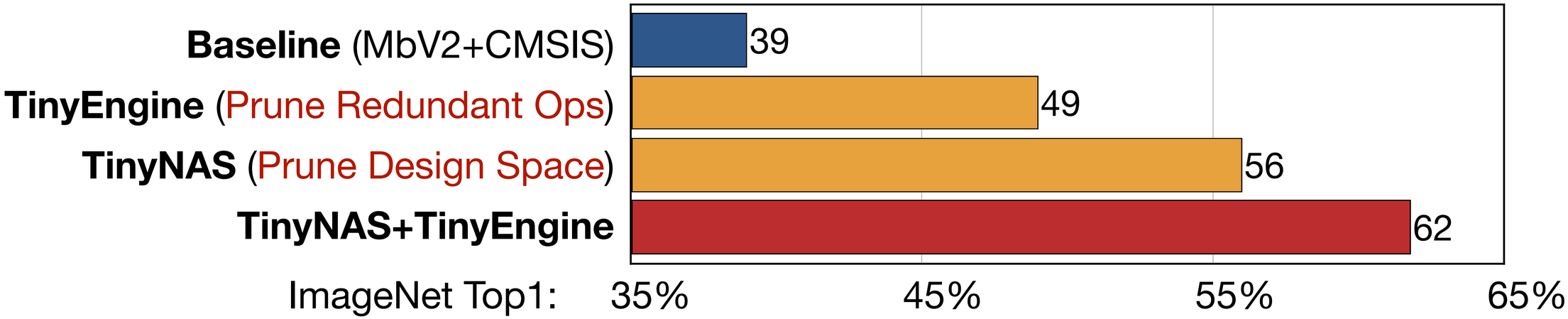


Song Han
MIT
songhan.mit.edu
mcunet.mit.edu

MCUNet: Tiny Deep Learning on IoT Devices

- Billions of IoT devices around the world based on microcontrollers (MCU).
- Low-cost ($1-2), low-power, small, almost everywhere in our lives.
- AI on MCU is **hard**: No DRAM. No OS. Extreme memory constraint.
- Existing work optimize for #parameters, but #activation is the **real bottleneck**.
- MCUNet: first to achieve >70% ImageNet top1 accuracy on a microcontroller.
- **Cloud AI**: ResNet; **Mobile AI**: MobileNet; **Tiny AI**: MCUNet. Demo.

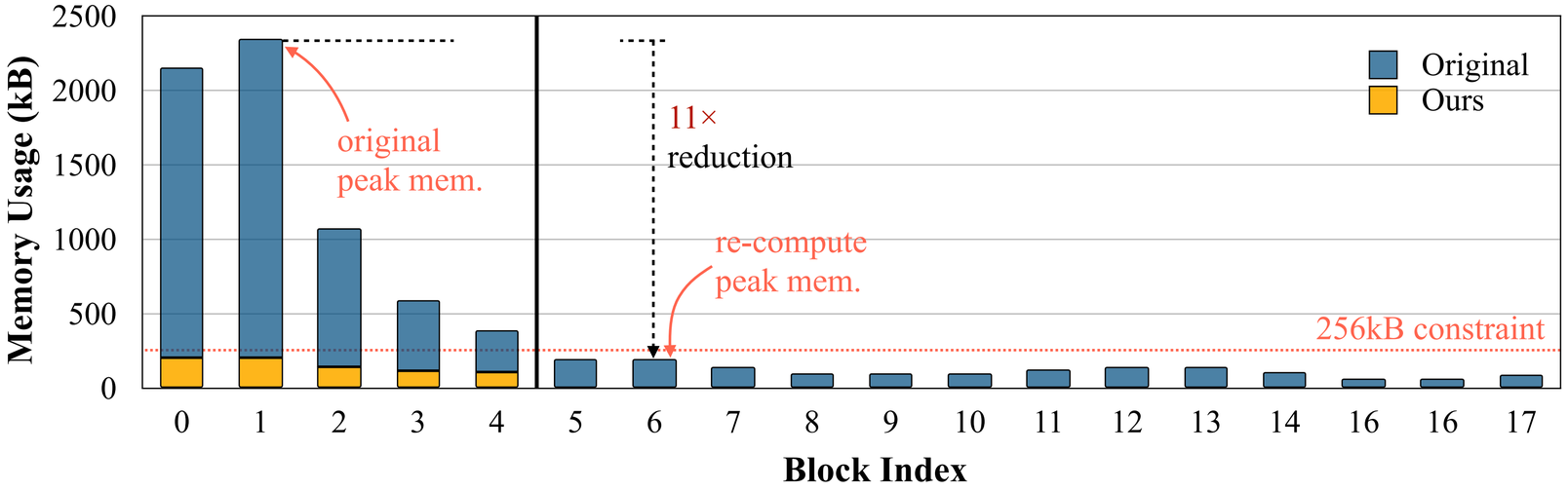


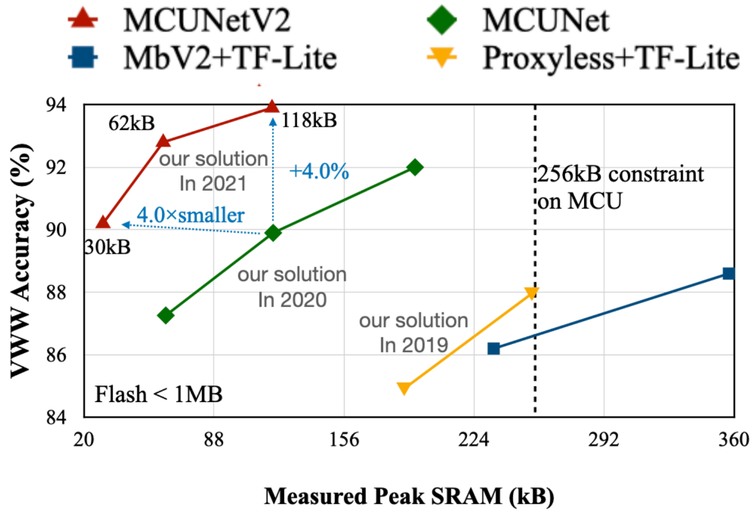


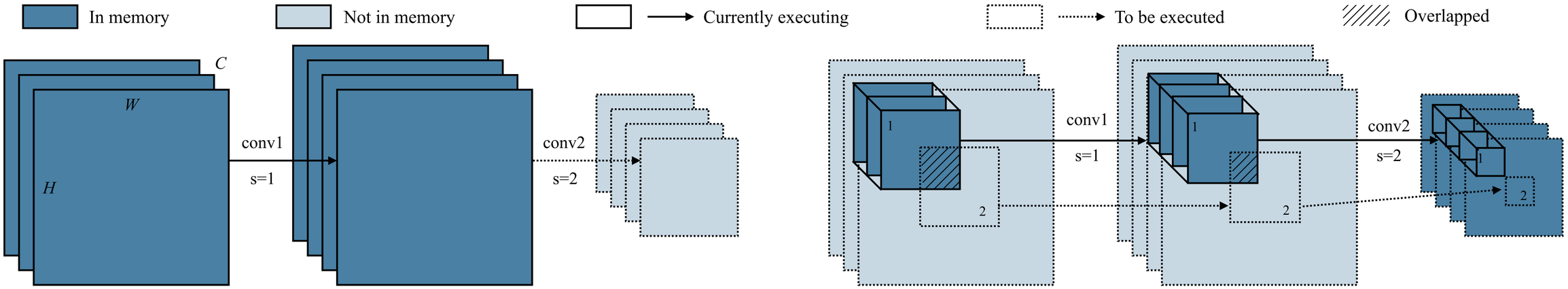
MCUNet, NeurIPS'20

MCUNet-v2: Memory-Efficient Patch-Based Inference

Problem: Imbalanced memory usage → activation bottleneck







(a) Per-layer computation (executing first conv)

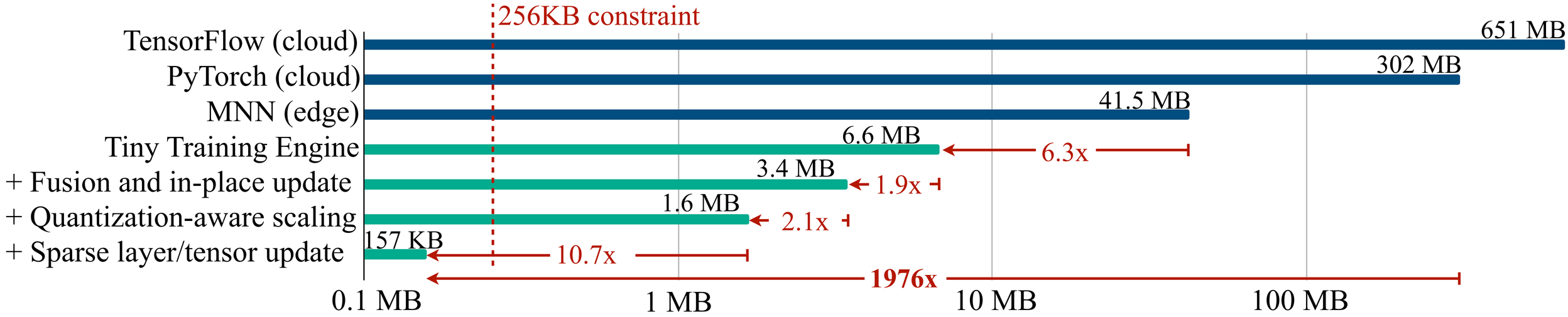
(b) Per-patch computation (executing first patch)

Demo: https://youtu.be/F4XKn0iDfxg

MCUNet V2, NeurIPS'21

MCUNet-v3: On-Device Training Under 256KB Memory

- AI systems need to adapt to **new** sensory data for **customization** and **continual learning.**
- Cloud-based learning leads to **privacy** issue and **high cost.**
- However, **training** is more expensive than **inference** due to back-propagation, making it hard to fit IoT devices (such as MCU only has 256KB SRAM).
- Idea: **sparse** layer / **sparse** tensor update + quantization-aware scaling on **real quantized graph** as opposed to fake quantized graph + tiny training engine: >1000x memory reduction.
- Demo



MCUNet-V3, Arxiv'22