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

MCUNetV3: On-Device Training Under 256KB Memory

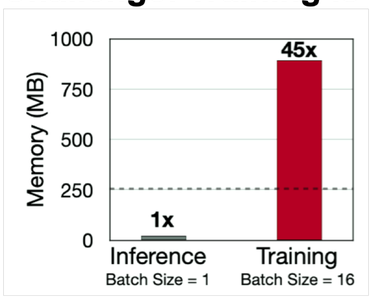
Ji Lin*, Ligeng Zhu*, Wei-Ming Chen, **Wei-Chen Wang**, Chuang Gan, Song Han
MIT, MIT-IBM Watson AI Lab

tinytraining.mit.edu

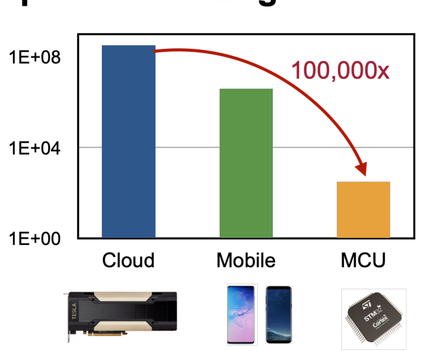
On-Device Training Enables Customization and Continual Learning

- **Security**: Data **never leaves devices,** thus promises security and regularization.
- **Customization**: Models **continually adapt** to new data from the sensors.

Challenge: Training is Expensive for Edge



(Measured with MobileNetV2)



Quantization-Aware Scaling (QAS)

- **Re-scale gradients** to help convergence with **quantized training**

Weight quantization

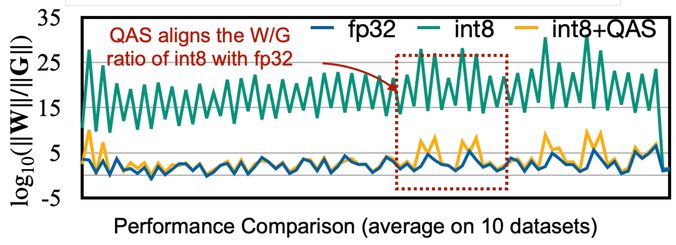
$$\mathbf{W} = s_{\mathbf{W}} \cdot (\mathbf{W}/s_{\mathbf{W}}) \overset{\text{quantize}}{\approx} s_{\mathbf{W}} \cdot \bar{\mathbf{W}}, \quad \mathbf{G}_{\bar{\mathbf{W}}} \approx s_{\mathbf{W}} \cdot \mathbf{G}_{\mathbf{W}},$$

Weight and gradient ratios are off by $S_{\mathbf{W}}^{-2}$

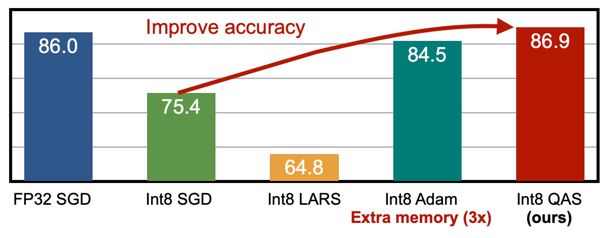
$$\|\bar{\mathbf{W}}\|/\|\mathbf{G}_{\bar{\mathbf{W}}}\| \approx \|\mathbf{W}/s_{\mathbf{W}}\|/\|s_{\mathbf{W}} \cdot \mathbf{G}_{\mathbf{W}}\| = \boxed{s_{\mathbf{W}}^{-2}} \cdot \|\mathbf{W}\|/\|\mathbf{G}\|.$$

Thus, re-scale the gradients

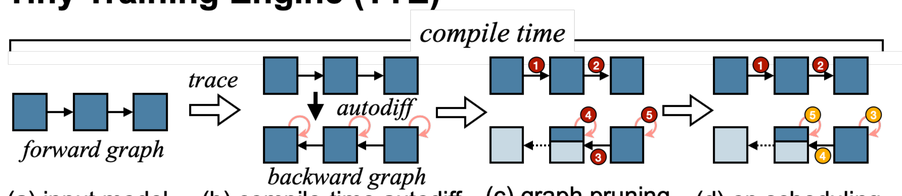
$$\tilde{\mathbf{G}}_{\bar{\mathbf{W}}} = \mathbf{G}_{\bar{\mathbf{W}}} \cdot s_{\mathbf{W}}^{-2}, \quad \tilde{\mathbf{G}}_{\bar{\mathbf{b}}} = \mathbf{G}_{\bar{\mathbf{b}}} \cdot s_{\mathbf{W}}^{-2} \cdot s_{\mathbf{x}}^{-2} = \mathbf{G}_{\bar{\mathbf{b}}} \cdot s^{-2}$$



Performance Comparison (average on 10 datasets)

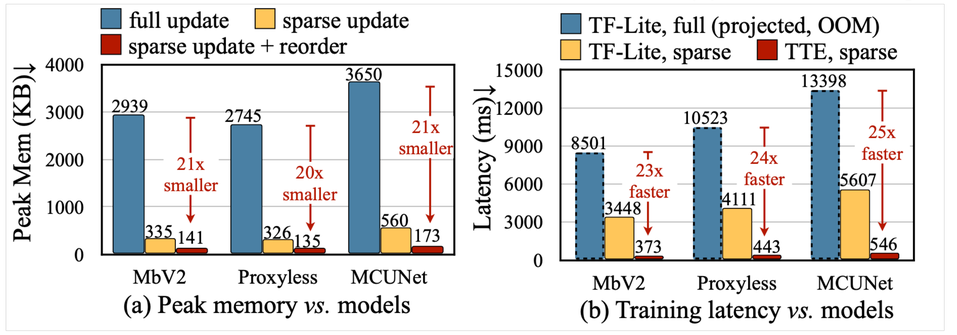


Tiny Training Engine (TTE)



Theoretical saving -> **Measured** saving:

- Offload most workloads from **runtime** to **compile-time**
- **Compile-time autodiff** enables more optimization opportunities.
- **Full support** for sparse update schemes

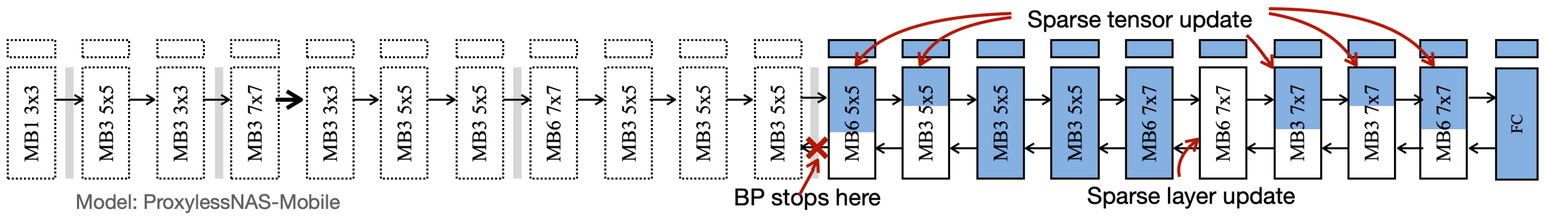


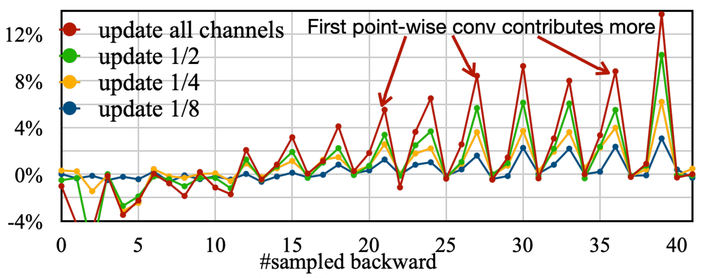
Sparse update saves **20~21x** memory

TTE kernels are **23~25x** faster.

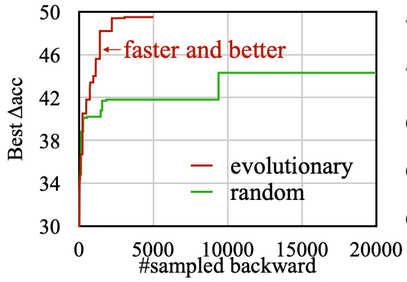
Sparse Layer/Tensor Update

- Conventionally, we update the **full model** for transfer learning
- We find some layers are **more important** than others, so we can **sparsely update**

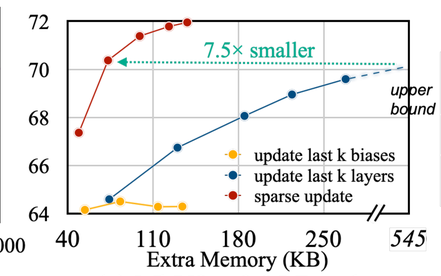




(a) Contribution analysis

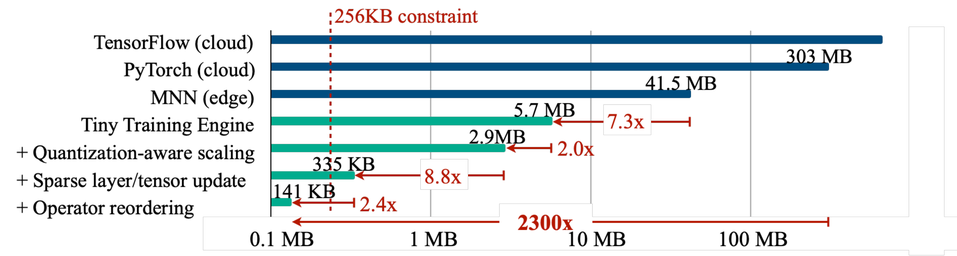


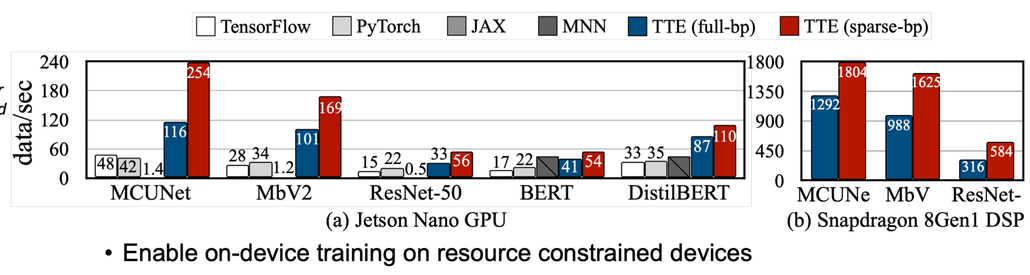
(a) Evolution Search



(c) MbV2-w0.35 Results

Algorithm-System Co-Design Results





- Enable on-device training on resource constrained devices
- 6x speedup compared to TensorFlow on Jetson

Acknowledgment to NSF Grant OAC-2117997