TinyML and Efficient Deep Learning make AI greener and deployable on IoT devices

Song Han
Assistant Professor
Massachusetts Institute of Technology
tinyml.mit.edu
Today’s AI is too Big

We need TinyML and Green AI

AlphaGo: 1920 CPUs and 280 GPUs, $3000 per game for electric bill
GPT-3: 175 billion parameters, 355 GPU years to train and cost $4.6M

Common carbon footprint benchmarks

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>CO2 Equivalent (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roundtrip flight b/w NY and SF (1 passenger)</td>
<td>1,984</td>
</tr>
<tr>
<td>Human life (avg. 1 year)</td>
<td>11,023</td>
</tr>
<tr>
<td>American life (avg. 1 year)</td>
<td>36,156</td>
</tr>
<tr>
<td>US car including fuel (avg. 1 lifetime)</td>
<td>126,000</td>
</tr>
<tr>
<td>Transformer (213M parameters) w/ neural architecture search</td>
<td>626,155</td>
</tr>
</tbody>
</table>

"Evolved Transformer with Neural Architecture Search" ICML’19, ACL’19
Deep Compression
compress an existing model by pruning & quantization

Pruning

Quantization

Original ResNet-50
with Deep Compression

100MB
6MB
17x compression

Deep Compression, ICLR'16, best paper award
Pruning & Sparsity

Increased attention since 2015

Optimal Brain Damage

Yann Le Cun, John S. Denker and Sara A. Solla
AT&T Bell Laboratories, Holmdel, N. J. 07733

Learning both Weights and Connections for Efficient Neural Networks

Song Han
Stanford University
songhan@stanford.edu

Jeff Pool
NVIDIA
jpool8@nvidia.com

John Tran
NVIDIA
johntran@nvidia.com

William J. Dally
Stanford University
NVIDIA
dally@stanford.edu

Han et al., NIPS’15
Can we go even smaller?

- The future belongs to Tiny AI.
- Billions of IoT devices around the world based on microcontrollers
- Much cheaper ($1-2), much smaller, almost everywhere in our lives.
- Low-cost: low-income people can have access. Democratize AI.
- Low-power: green AI, reduce carbon

finecontrol

Smart Home

Smart Manufacturing

Personalized Healthcare

Driving Assist
tinyml.mit.edu
But it’s challenging:

- Tiny model design is fundamentally different.
- No DRAM. No operating system (no virtual memory).
- Extreme memory constraint. Even for the code segment. No space for interpreter.
- Can’t directly scale. Existing work optimize for #parameters, but #activation is the real bottleneck.

<table>
<thead>
<tr>
<th></th>
<th>Cloud AI</th>
<th>Mobile AI</th>
<th>Tiny AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>32GB</td>
<td>4GB</td>
<td>256kB</td>
</tr>
<tr>
<td>Storage</td>
<td>~TB/PB</td>
<td>256GB</td>
<td>1MB</td>
</tr>
</tbody>
</table>

16,000x smaller
100,000x smaller

tinyml.mit.edu
MCUNet: Bring AI to IoT Devices

MCUNet achieves >70% ImageNet accuracy with only 512KB of memory

• Running on STM32 MCU, Cortex-M7 @216MHz
• 320KB SRAM, 1MB Flash
• Highlighted by MIT News, Wired, NSF Newsletter, IBM

Without GPU or any specialized HW, MCUNet is so tiny that it can run on IoT device.

[NeurIPS’20]
[NeurIPS’21]
MCUNet: Bring AI to IoT Devices

Detect person using only 30KB of memory!
Unlock ultra low-power AIoT Applications

![Graph showing performance vs. memory usage for different models](chart.png)

- MCUNet-v2 [NeurIPS'21]
- MCUNet [NeurIPS'20]
- MbV2+TF-Lite
- Proxless+TF-Lite

- VVW Accuracy (%)
- Measured Peak SRAM (kB)
- Flash < 1MB
- 62kB - 118kB
- 4.0× smaller
- +4.0%
- 256kB constraint on MCU
TinyML for Point Cloud & LiDAR Processing

- 3D point cloud models: 10x more computationally expensive than 2D CNNs
- Challenge: highly sparse & irregular, large memory footprint
- Random memory access is unfriendly for CPU/GPU/TPU => customized system & HW

![Diagram of point cloud processing algorithms and hardware systems]

New design space, new primitive for point cloud

**Algorithm**

3D neural architecture search

**Hardware**

[PointAcc, MICRO’21]

Hardware accelerator for point cloud

[Point-Voxel CNN, NeurIPS’19 spotlight]

**System**

[SPVNAS, ECCV’20]

[SPVNAS, ECCV’20]

Automotive VR AR

3D neural architecture search

[GPU#1]

[TorchSparse, open source]

GPU library for 3D sparse convolution

[PointAcc, MICRO’21]
OFA Designs Light-weight Model, Bring AI to Mobile Devices

Running on LG phone with Qualcomm Snapdragon 855 SoC (released in 2019)
Software solution to achieve real-time inference on mobile device even without AI accelerator
Unlock many mobile AI applications: healthcare, smart home, automotive…

- on-device car/person detection
- on-device pose estimation
- on-device segmentation
Anycost GAN

- Generative Adversarial Network (GAN) is computationally heavy and slow
- Difficult for interactive photo editing on mobile device (iPad)
- Anycost GAN with once-for-all network:
  - Small sub-net: low cost, fast prototyping
  - Large sub-net: high-quality finalization

Running on 2019 MacBook
Industry Integration

**Once-for-All (OFA) Network** integrated by Alibaba received a world-record in the open division of MLPerf Inference Benchmark, achieving 1.078M images per second on eight A100 GPUs.

**Once-for-All (OFA) Network** integrated by Maxim Integrated provides 6% accuracy increase in image recognition and 2% accuracy increase in speech command recognition, with >100x energy efficiency compared to Cortex-M4.

**Proxyless Neural Architecture Search**, an efficient neural architecture search algorithm with light-weight model for mobile AI is integrated by Amazon AutoGluon and Facebook PyTorch.

- First place, 6th AI Driving Olympics, NuScenes Segmentation Challenge @ICRA’21
- First place, 5th Low-Power Computer Vision Challenge, CPU detection track & FPGA track
- First place, 3D semantic segmentation challenge on SemanticKitti
- First place, 4th Low-Power Computer Vision Challenge, CPU classification and detection track
- First place, 3rd Low-Power Computer Vision Challenge, DSP track, @ICCV’19
- First place, MicroNet Challenge, NLP track (WikiText-103), @NeurIPS’19
- First place, Visual Wake Words Challenge, TF-lite track, @CVPR’19