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

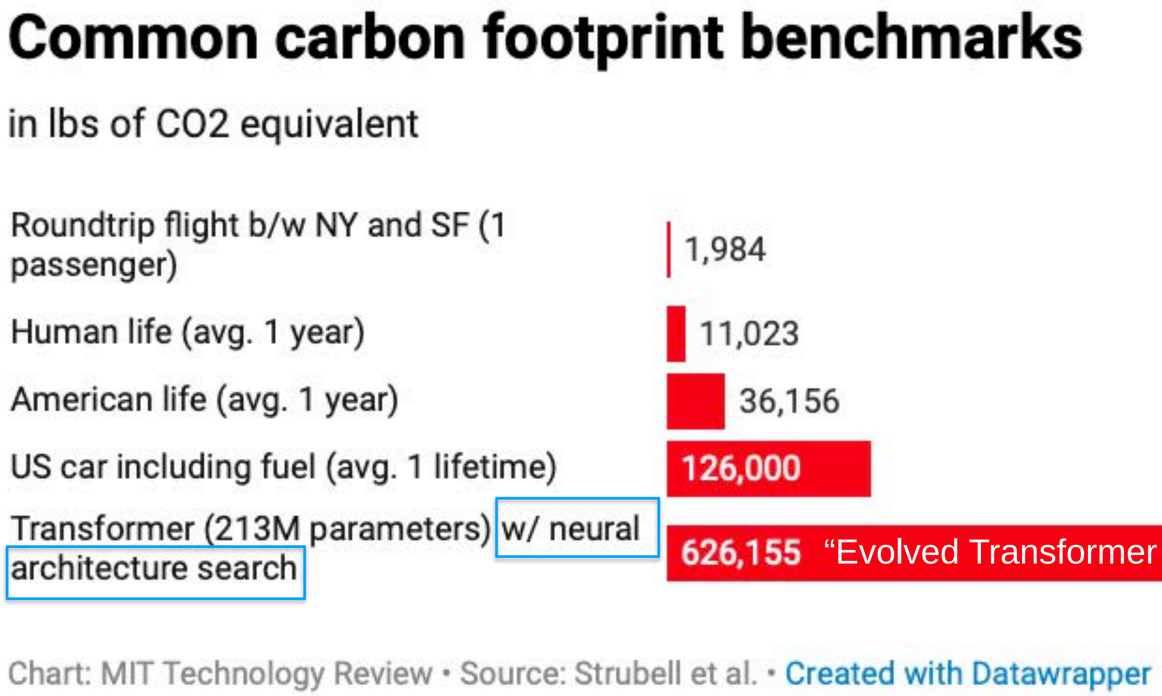
AlphaGo: 1920 CPUs and 280 GPUs, $3000 per game for electric bill
GPT-3: 175 billion parameters, 355 GPU years to train and cost $4.6M

Artificial intelligence / Machine learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by **Karen Hao** June 6, 2019

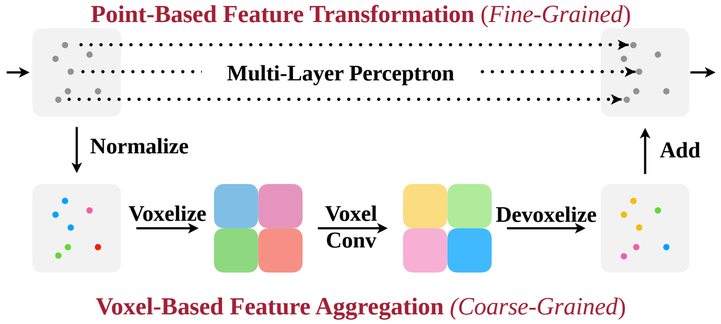


"Evolved Transformer with Neural Architecture Search" ICML'19, ACL'19

TinyML for Point Cloud & LiDAR Processing

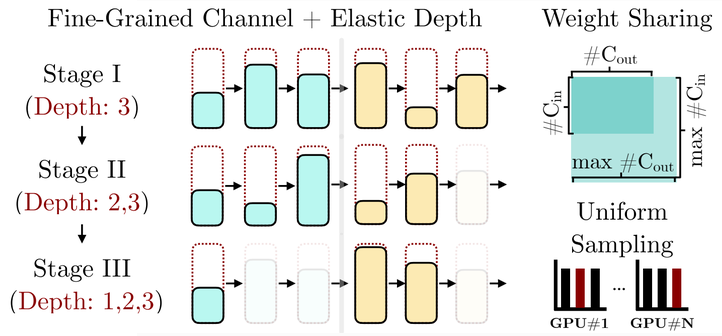
[PVCNN, NeurIPS'19]
[SPVNAS, ECCV'20]
[PointAcc, MICRO'21]

- 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



[Point-Voxel CNN, NeurIPS'19 spotlight]

New design space, new primitive for point cloud

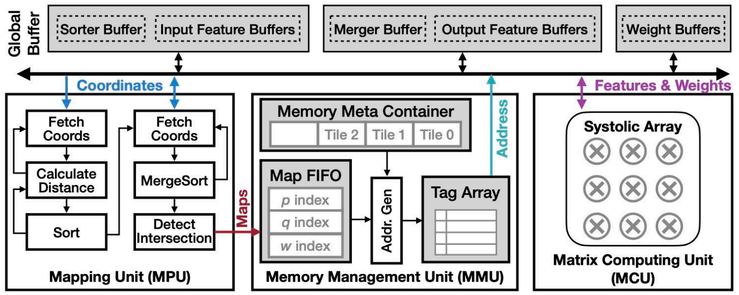


[SPVNAS, ECCV'20]

3D neural architecture search

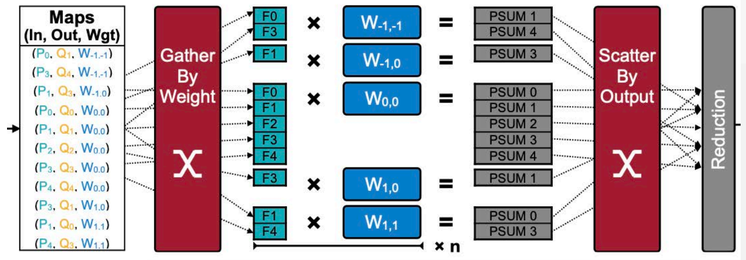
Algorithm

Hardware | **System**



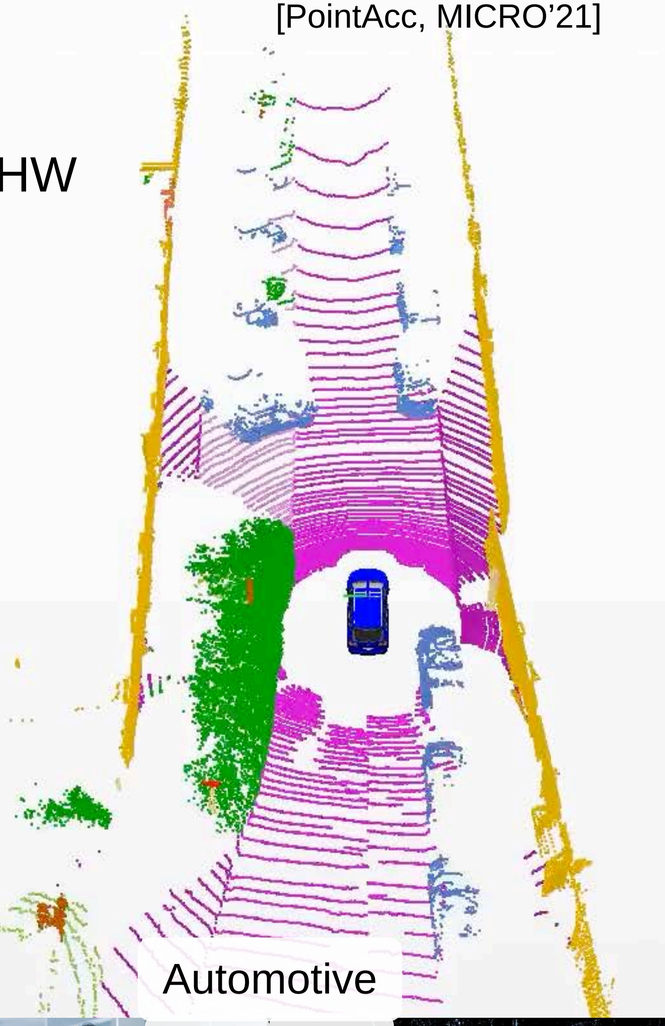
[PointAcc, MICRO'21]

Hardware accelerator for point cloud



[TorchSparse, open source]

GPU library for 3D sparse convolution



Automotive

TinyML for Point Cloud & LiDAR Processing

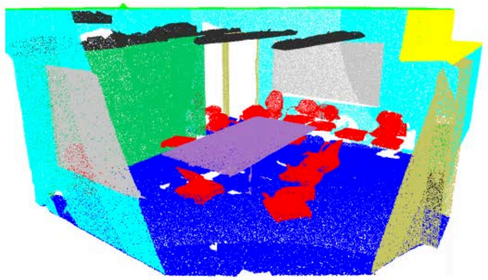
[PVCNN, NeurIPS'19]
[SPVNAS, ECCV'20]
[PointAcc, MICRO'21]

Ranks 1st in the nuScenes LiDAR Segmentation Challenge
Best submission@6th AI Driving Olympics, ICRA 2021

Object Part Segmentation

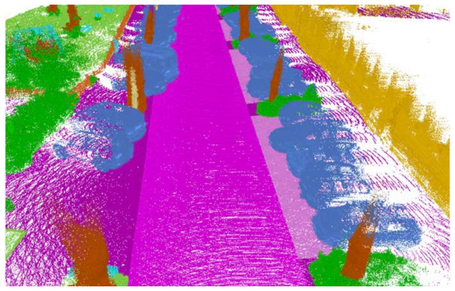
2.7x measured speedup

Indoor Scene Semantic Segmentation



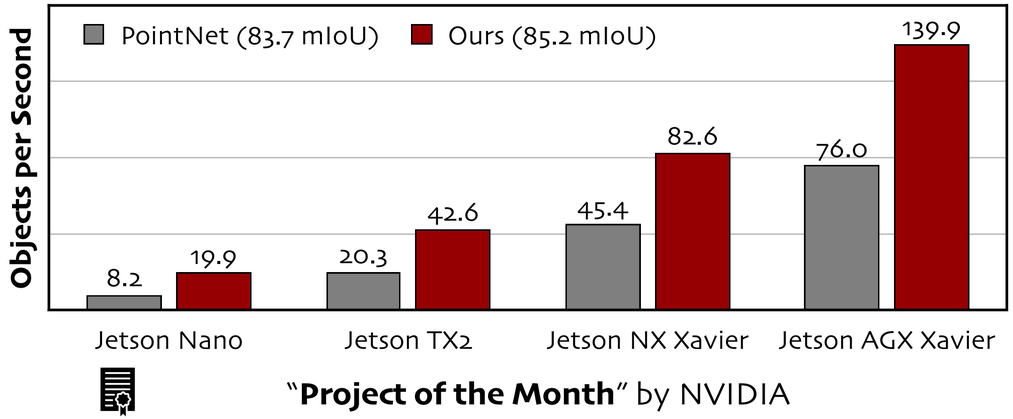
6.9x measured speedup

Outdoor Scene Semantic Segmentation

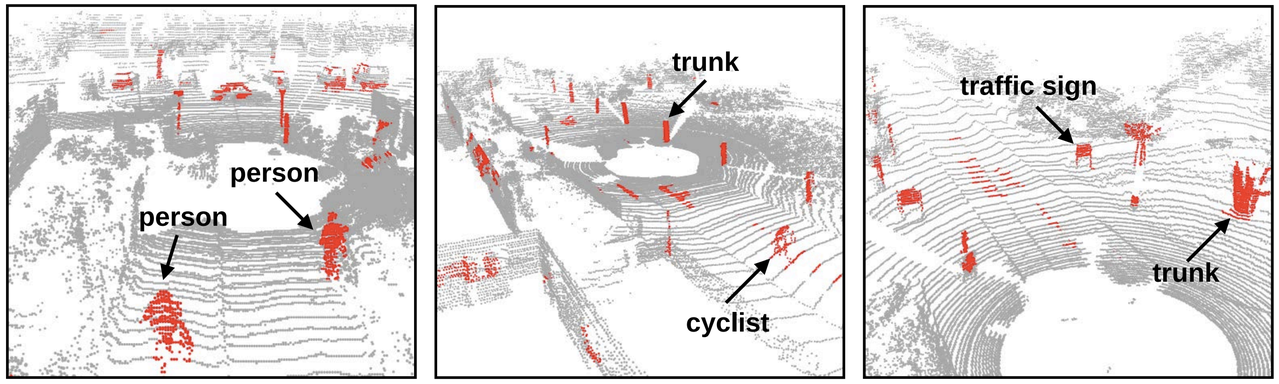


2.7x measured speedup

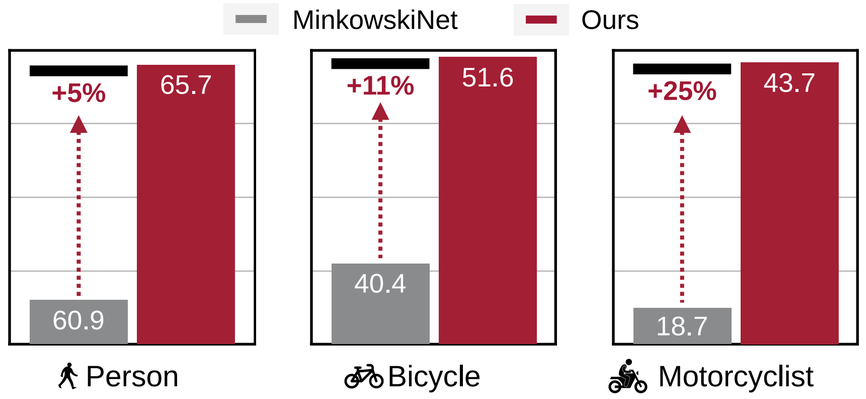
Real-Time Inference, 2x speedup on Edge Devices



"**Project of the Month**" by NVIDIA



PVCNN provides fine details for small objects at low computation



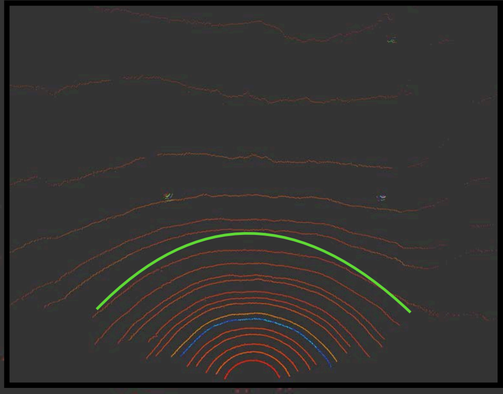
Significant accuracy improvement on safety-critical objects

MIT Driverless

Accuracy: **95.0%**

Range: **8 meters**

Latency: **2 ms/object**

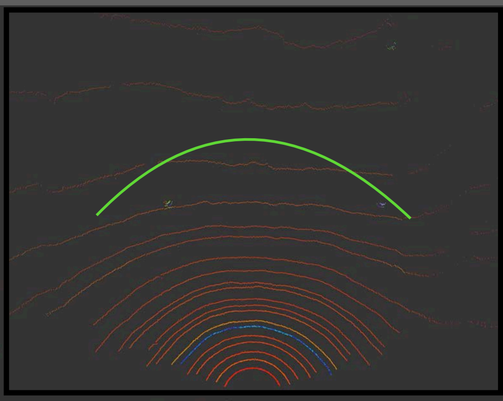


PVCNN (Ours)

Accuracy: **99%**

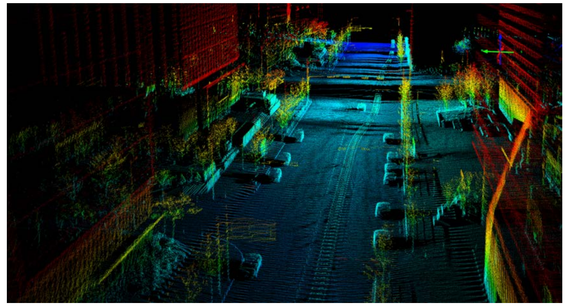
Range: **12 meters**

Latency: **1.25 ms/object**

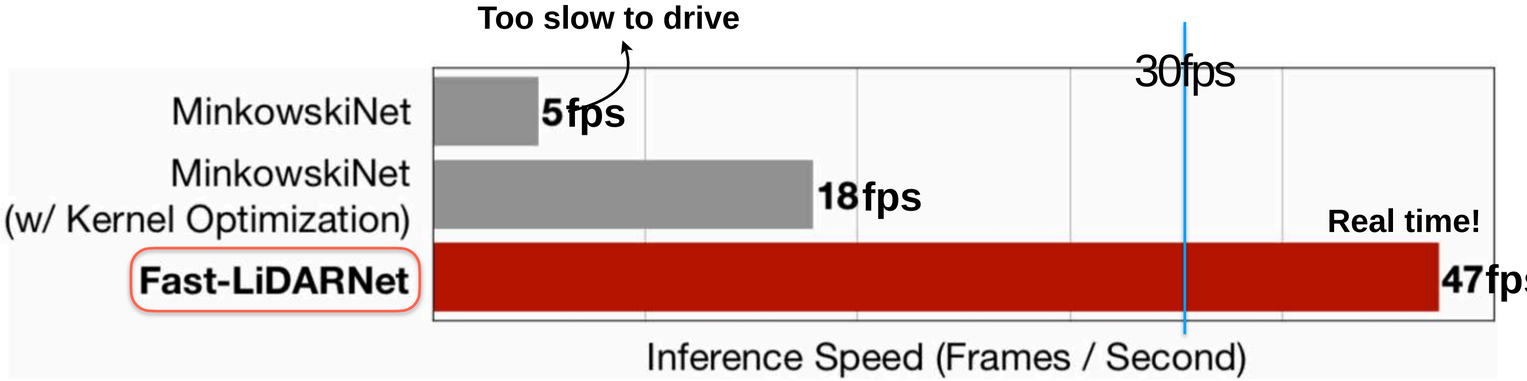


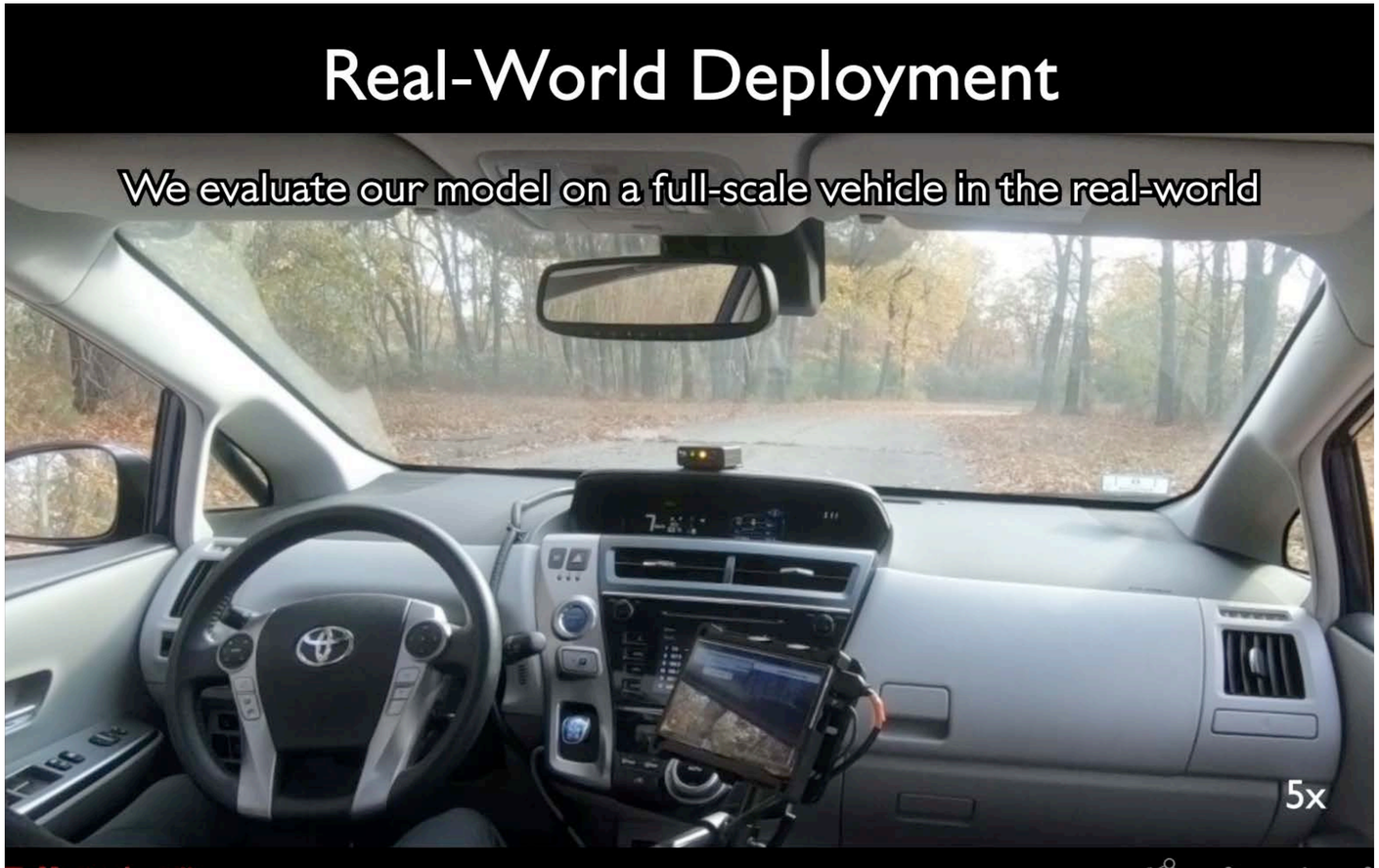
TinyML for Point Cloud & LiDAR Processing

3D LiDAR Sensor



3D Point Cloud: 2M points/s





[Liu et al. ICRA'21]

In collaboration
with Daniela Rus