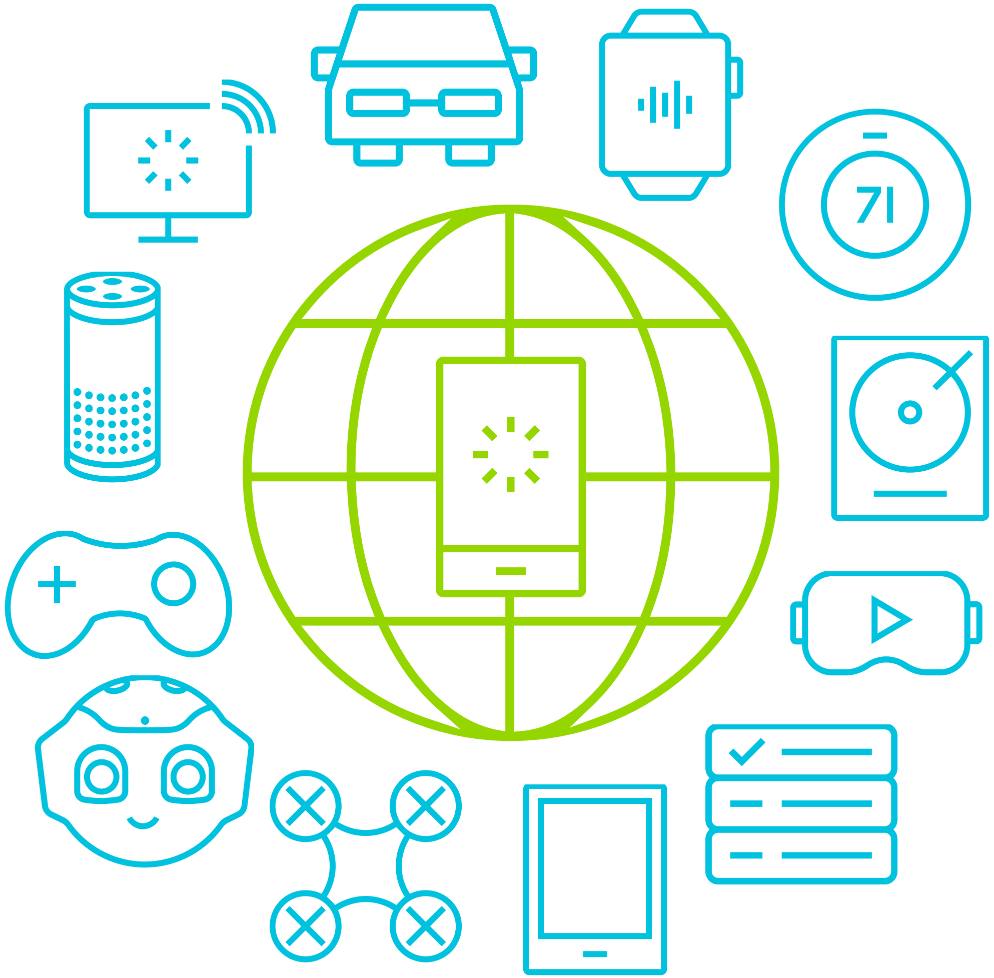
Arm ML Research Project Overview

Mark O'Connor
Senior Principal Researcher
Arm Machine Learning Research Lab

Arm Provides Compute

#1
shipping GPU in the world is Mali

23Bn
Arm-based chips shipped in 2018



> 5Bn
people using Arm-based mobile phones

146Bn
Arm-based chips to date

Neural Networks as Software 2.0

A **fundamental shift** in how we write software

Program is **learned** rather than written

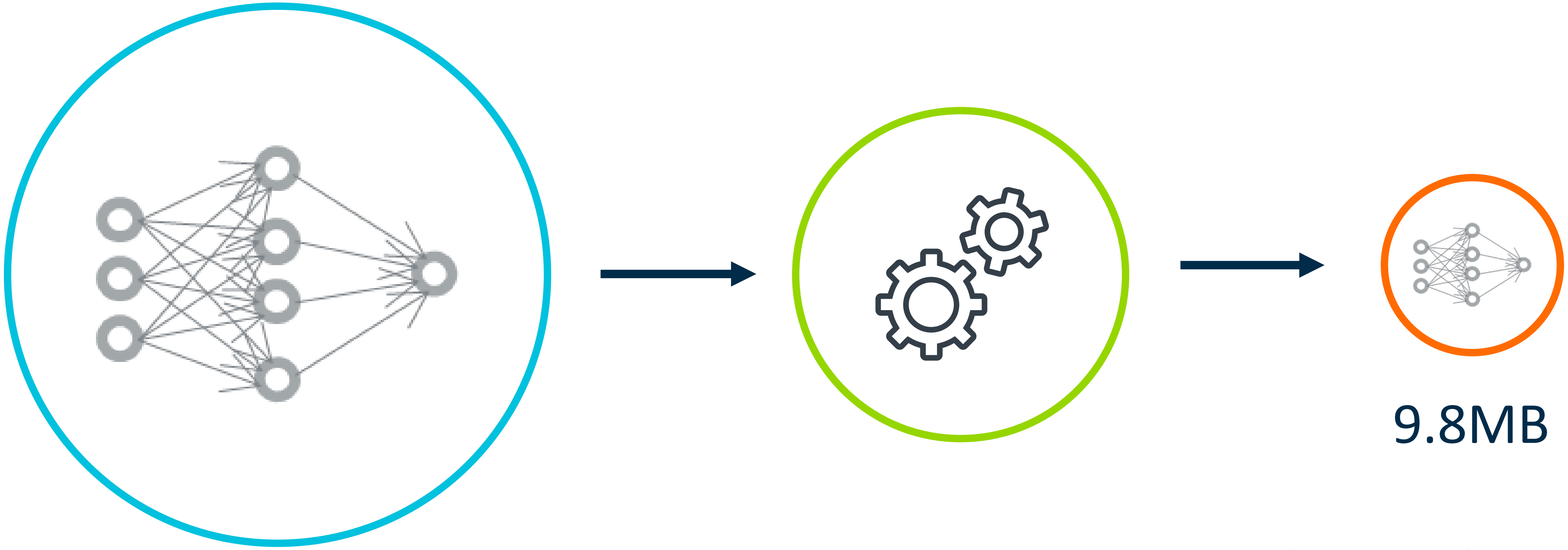
Training compiles **data** directly into **weights**

Weights execute as a **computational graph**

Andrej Karpathy
Director of AI at Tesla

Pete Warden
TensorFlow Engineer at Google

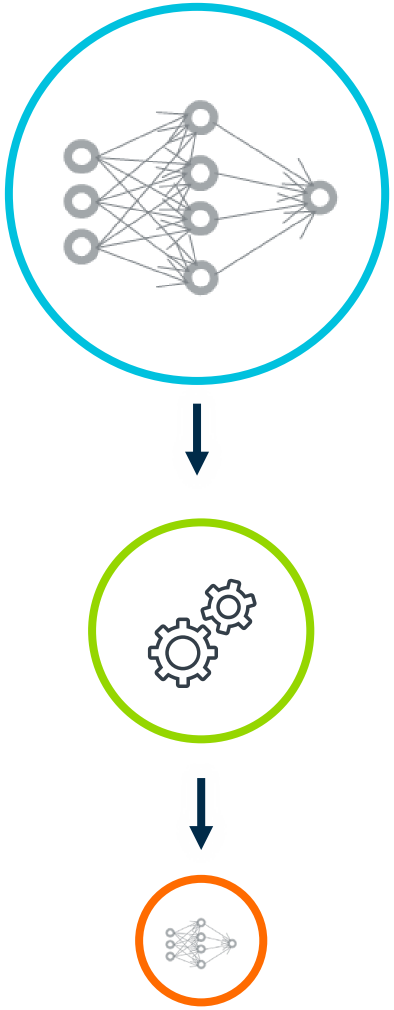
Software 2.0 is Surprisingly Adaptable



95.7MB

10x reduction in model size with
1% accuracy loss on Inception v3

Achieving Trillions of Operations-per-Second on a Mobile Platform



Reduced precision	FP32 to INT8	4x throughput
Data re-use	Exploit reuse weights	Convolution filters
Weight compression	Reduce memory footprint	Better bandwidth and power
Transforms	Winograd convolutions	$O(n^3) \Rightarrow O(n^2)$
Sparse computation	ReLU activation	Pruning techniques

TinyML Hardware Trends

- Low-power general purpose cores and SoCs
  - Wide (vector) registers to support machine learning workloads
  - Dot-product and matrix multiply ops
- NPU
  - Hardware support for lower precision (<=4b) operations
  - Hardware support for sparsity/zero-skipping
  - Hardware support for compression of weights and activations
- Analog & Compute-in-memory
  - First commercial products appearing

Amortize front-end overhead

Key operation in all neural networks

Recent research shows minimal accuracy loss with 4b

ReLU activation and weight sparsity

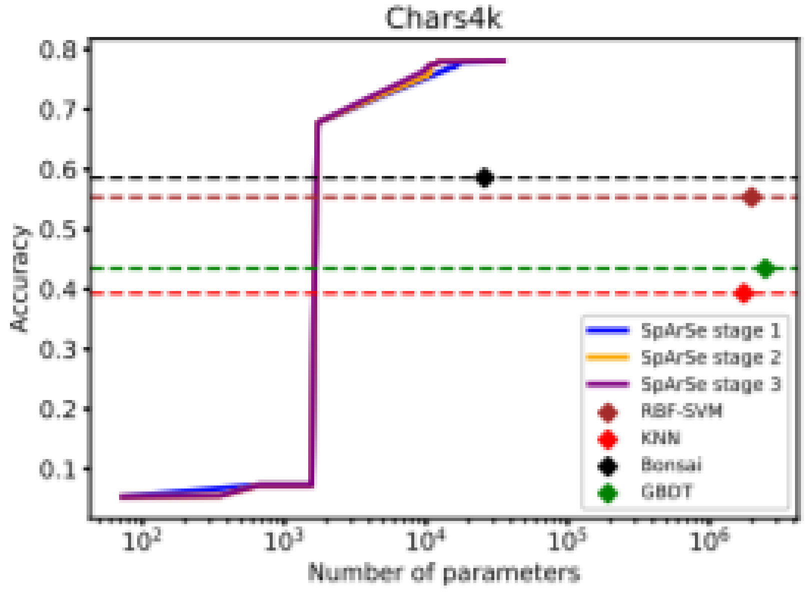
Bandwidth bottleneck, esp. FC

arm

SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers

- Microcontrollers are small, ubiquitous processors:
  - At the heart of the Internet of Things
  - Limited RAM and ROM – as little as 10s of KB
- We want to find CNN designs that:
  - Fit on microcontrollers
  - Deliver state-of-the-art accuracy
- The SpArSe framework:
  - Combines network architecture search and network pruning
  - Multi-objective Bayesian optimizer generates configurations
  - Finds highly accurate neural net models with significantly fewer parameters
  - Presented at NeurIPS 2019 in Vancouver, Canada

Processor	Usecase	Compute	Memory	Power	Cost
Nvidia 1080Ti GPU [3]	Desktop	10 TFLOPs/Sec	11 GB	250 W	$700
Intel i9-9900K CPU [6, 5]	Desktop	500 GFLOPs/Sec	256 GB	95 W	$499
Google Pixel 1 (Arm CPU) [10]	Mobile	50 GOPs/Sec	4 GB	~ 5 W	–
Raspberry Pi (Arm CPU) [11]	Hobbyist	50 GOPs/Sec	1 GB	1.5 W	–
Micro:Bit (Arm MCU) [8]	IoT	16 MOPs/Sec	16 KB	~ 1 mW	$1.75
Arduino Uno (Microchip MCU) [1]	IoT	4 MOPs/Sec	2 KB	~ 1 mW	$1.14

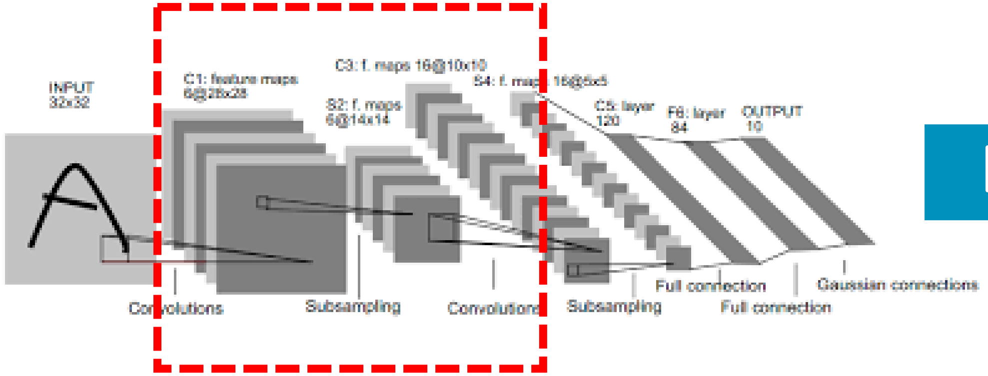


https://arxiv.org/abs/1905.12107

How Can We Improve Performance/Power/Area Further?

DeepFreeze: A hardware-generating backend for TensorFlow

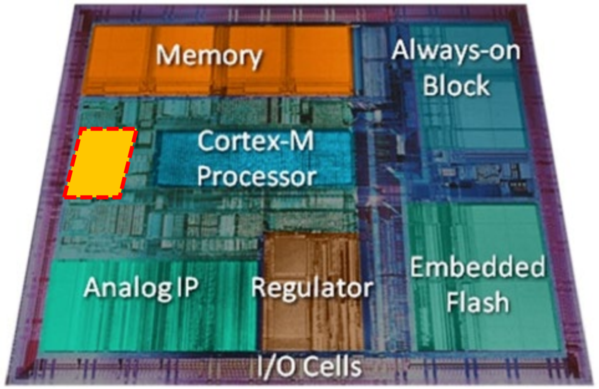
NN model in TensorFlow



Specified layers or whole model

DeepFreeze

Hardware accelerator in Verilog



ASIC

FPGA

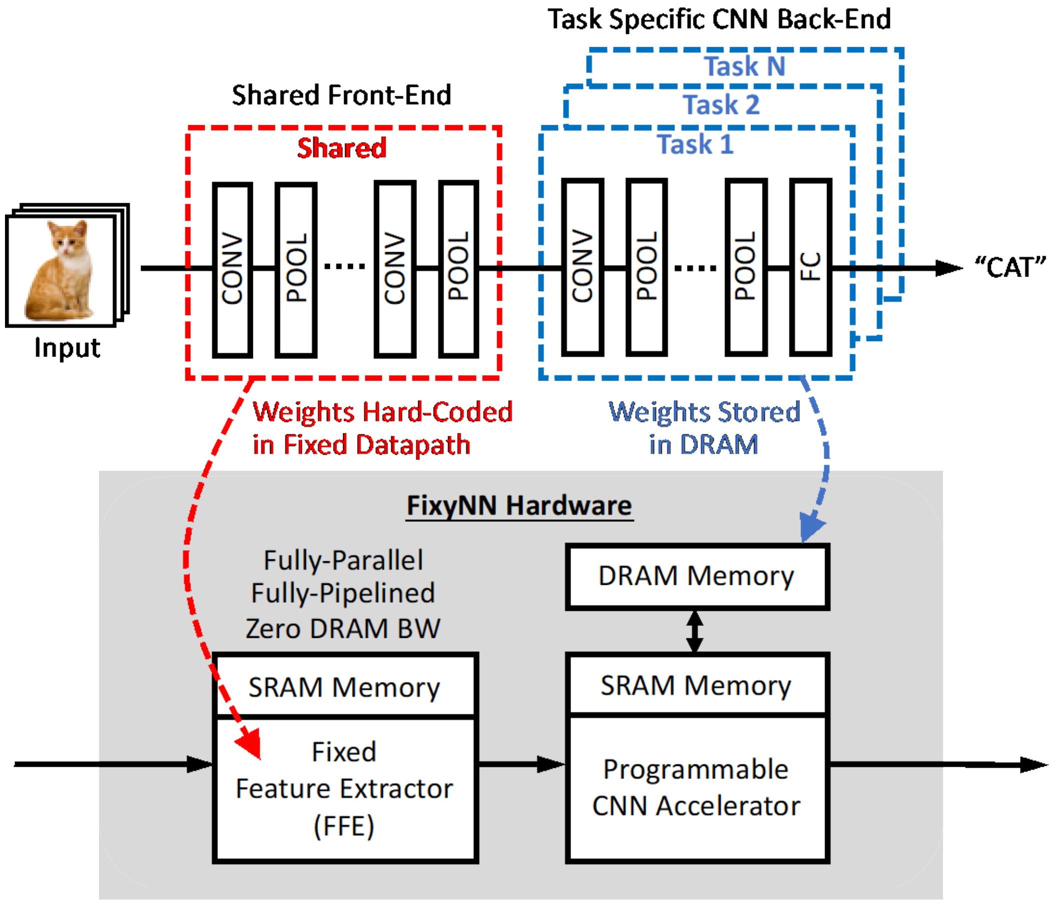
https://github.com/ARM-software/DeepFreeze

arm

How Can We Improve Performance/Power/Area Further?

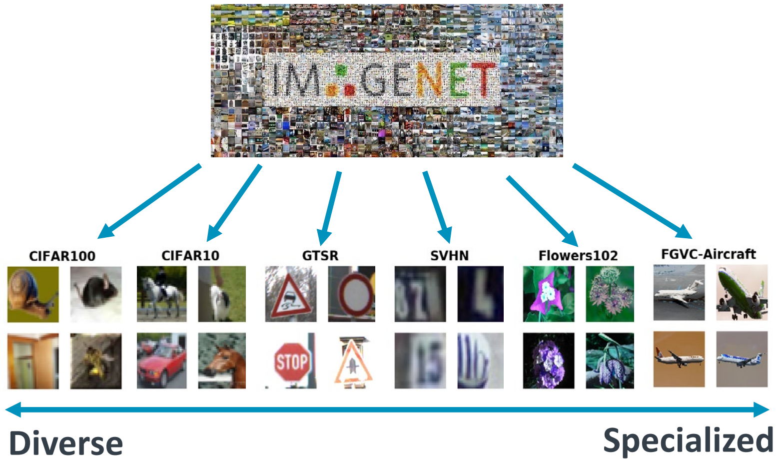
FixyNN: Leveraging Transfer Learning for Ultra-Efficient Hardware

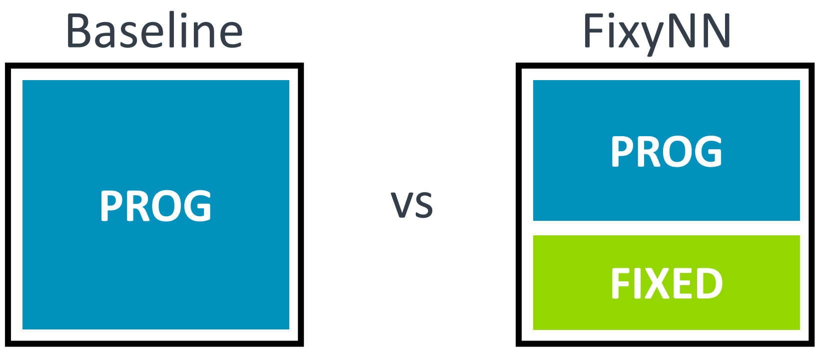
Train	• Perception architecture such as MobileNet on a wide range of images
Fix	• Implement first K layers in silicon with fixed weights
Tune	• Remaining N-K network layers for specific task
Run	• Task-specific part on programmable accelerator
Win	• Higher throughput and lower power compared to a fully-programmable accelerator

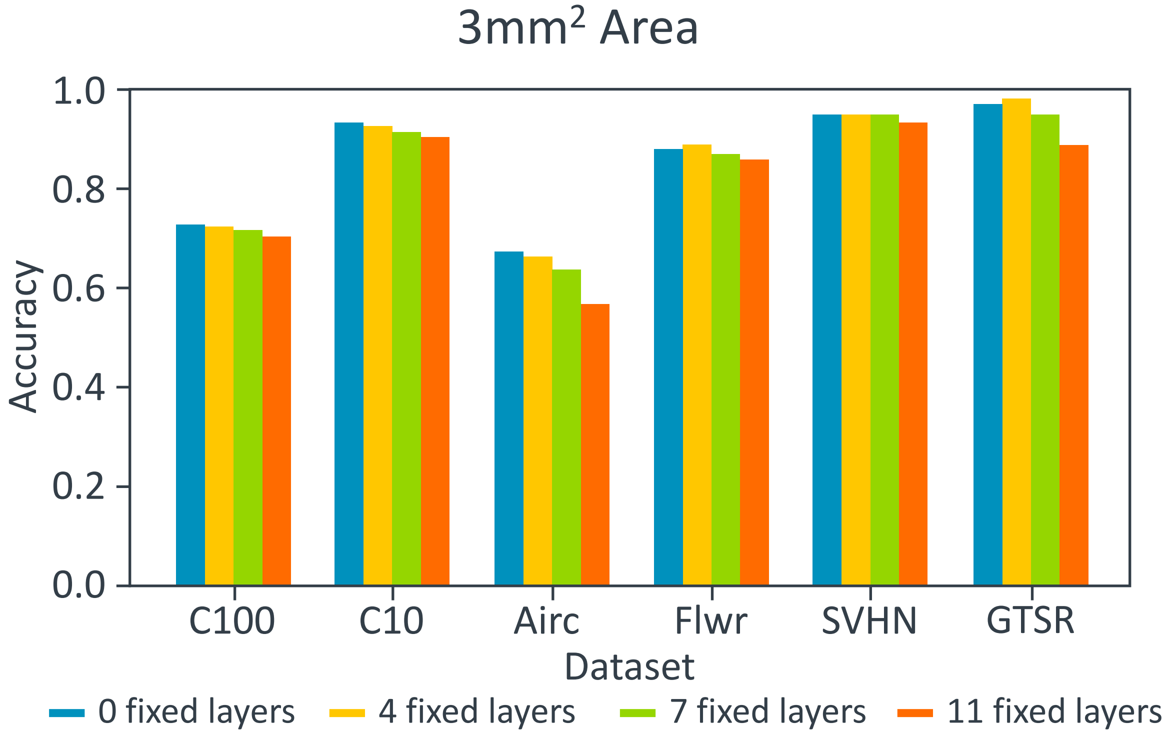


FixyNN Evaluation Results

- Energy efficiency of up to **11.2 TOPS/W** with <1% accuracy loss – nearly **2 ×** **more efficient** than NVDLA alone in same area
  - **~1.5x** TOPS/W by fixing 4 layers
  - **~2x** TOPS/W by fixing 7 layers
  - Accuracy loss of < 1% over six datasets







P. Whatmough, **C. Zhou**, **P. Hansen**, S. Venkataramanaiah, J. Seo, **M. Mattina**, "FixyNN: Efficient Hardware for Mobile Computer Vision via Transfer Learning," 2019 Conference on Systems and Machine Learning (SysML '19)

arm

Recent Publications from Arm ML Research Lab

arm

Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
ধন্যবাদ
תודה