arm

Machine Learning Research at Arm

Matthew Mattina

Senior Director, Arm ML Research Lab

Arm ML Research Lab: Vision and Core Research Threads

Arm's ML Research Lab vision is to be a technology leader in efficient ML inference and distributed ML

Efficient Hardware for ML

- Hardware designs and accelerator microarchitecture
- ISA additions
- Exploiting emerging device technology

Model Design & Optimization

- Novel Model architectures
- Emerging Use cases
- AutoML and Network Architecture Search

Edge-to-Cloud & ML Systems

- Distributed and on-device training
- Model security and model distribution

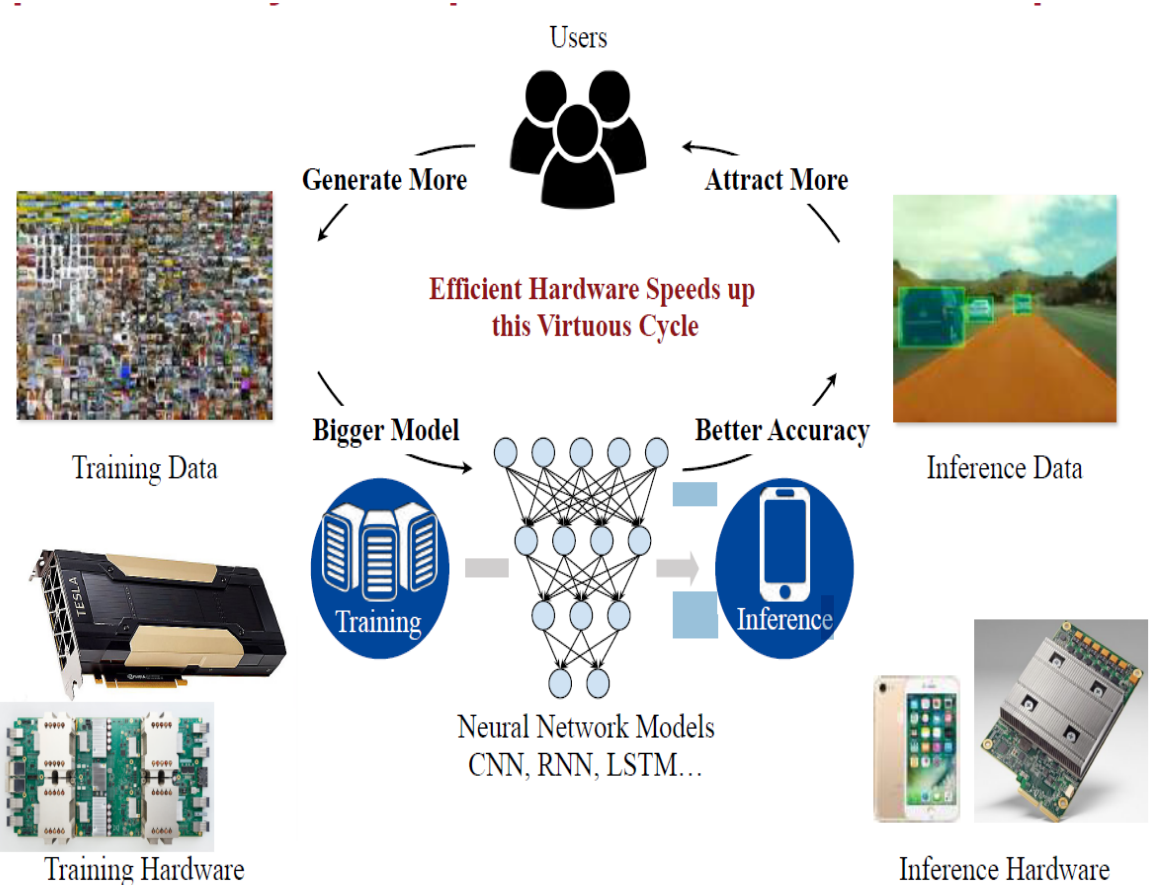
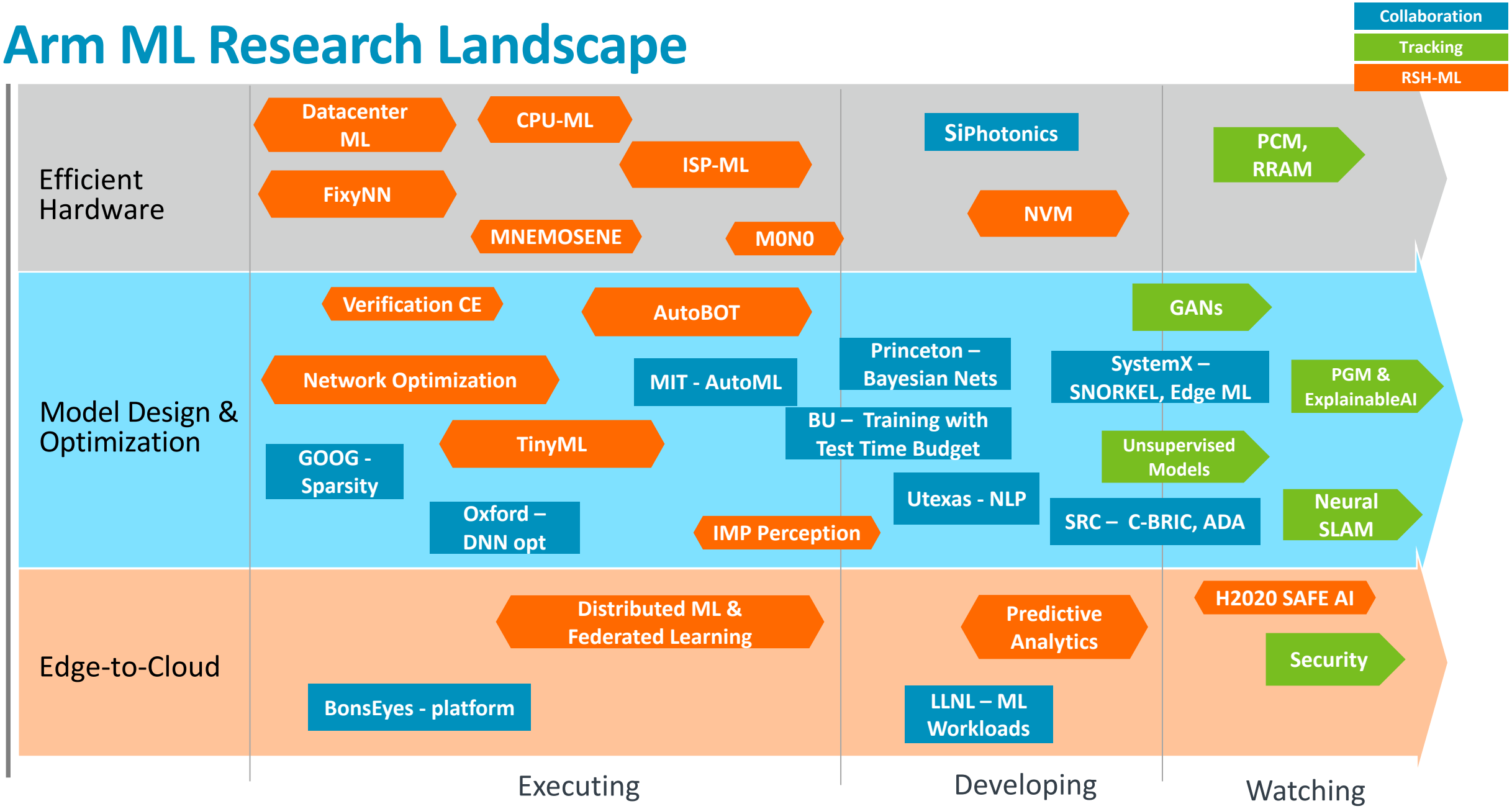


Image credit: Song Han

Arm ML Research Landscape



arm

Arm ML Research Lab


Matthew Mattina


Andrew Mundy


Hokchhay Tann


Chu Zhou


Dibakar Gope


Partha Maji


Igor Fedorov


Jesse Beu


John Brown


Patrick Hansen


Ramon Matas


Urmish Thakker


Paul Whatmough


Zhi-Gang Liu

We’re Hiring!

Openings available in our Boston, Austin, and Cambridge UK Locations!

arm

Recent Publications from Arm ML Research Lab

I. Fedorov, R. Adams**, M. Mattina, P. Whatmough** “SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers” (NeurIPS ‘19)

Zhi-Gang Liu, M. Mattina, “Learning low-precision neural networks without Straight-Through Estimator(STE)” (IJCAI ‘19)

D. Gope, G. Dasika, **M. Mattina**, “Ternary Hybrid Neural-Tree Networks for Highly Constrained IoT Applications,” 2019 Conference on Systems and Machine Learning (SysML ‘19)

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)

U. Thakker, **J. Beu**, G. Dasika, **M. Mattina**, “Measuring scheduling efficiency of RNNs for NLP Applications,” International Workshop on Performance Analysis of Machine Learning Systems (FastPath ’19)

U. Thakker, **J. Beu**, **D. Gope**, G. Dasika, **M. Mattina**, “RNN Compression using Hybrid Matrix Decomposition,” (tinyML Summit ’19)

P. Maji, A. Mundy, **G. Dasika**, **J. Beu**, **M. Mattina**, R. Mullins, “Efficient Winograd or Cook-Toom Convolution Kernel Implementation on Widely Used Mobile CPUs,” Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC^2 ‘19)

P. Whatmough, C. Zhou, P. Hansen, M. Mattina, “Energy Efficient Hardware for On-Device CNN Inference via Transfer Learning”, On-Device ML Workshop, Neural Information Processing Systems (NeurIPS ‘18)

Y. Zhu, A. Samajdar, **M. Mattina**, **P. Whatmough**, “Euphrates: Algorithm-SoC Co-Design for Low-Power Mobile Continuous Vision”, International Symposium on Computer Architecture (ISCA’18)

University Engagements

University	Topics	Status/Agreement
Harvard University	Sasha Rush, David Brooks, GuYeon Wei NLP NN/HW co-design	Funded by Arm RSH-ML over three years, 2018-2020
MIT	HAN Lab - Song Han Deep Compression AutoML	Funded by Arm RSH-ML over three years, 2018-2020
Boston University	Venkatesh Saligrama Learning for a test-time Budget Learning with Limited Supervision	Funded by Arm RSH-ML over three years, 2018-2020
Princeton University	Ryan Adams Co-optimization of ML / hardware Simple, robust, decision-making machine	Funded by Arm RSH-ML over three years, 2018-2020
Trinity College Dublin	CALCULUS - performance optimization techniques	Funded by Arm over four years, 2017-2020
Oxford University	Nic Lane Binary network optimization Statistic foundation for network pruning	Funded by iCASE over three years 2017-2019 PhD student, Javier Fernández-Marqués, to intern in 2019
SRC/GRC	JUMP Center liaison: CBRIC-T1: Neuro-inspired Algorithms & Theory JUMP Center tracking: ADA Hadi Esmaeilzadeh, UCSD, cloud-to-edge stack for DNN acceleration	Funded by Arm RSH
SystemX, Stanford	Computation for Data Analytics - Chris Re SNORKEL, ML for the Edge	Funded by Arm RSH for two Focus Area tokens
RISElabs, Berkeley	Data services vision Clipper - a general-purpose ML model serving system Ray - a distributed execution framework - cloud-edge ML	Funded primarily by ISG plus RSH & IPG contributions
BonsEyes	ARM-based platforms as example deployments	Funded by EU over three years, 2017-2019
University of Texas, Austin	Greg Durrett - NLP, scalable training/inference with large data sets Ruben Rathnasingham - Collaboration with Dell Medical	Funded by Arm RSH for 2019 On standby for consulting
University of Michigan	Honglak Lee - Generative networks, GANs	Arm RSH - UoM sponsorship, exploring topics for collaboration
University of Manchester	Gavin Brown - Ensembles to Modular NNs	Exploring topics for collaboration
University of Cambridge	TBD	CDT program, exploring topics for collaboration

arm

Arm ML Research Lab: Selected Projects

TinyML

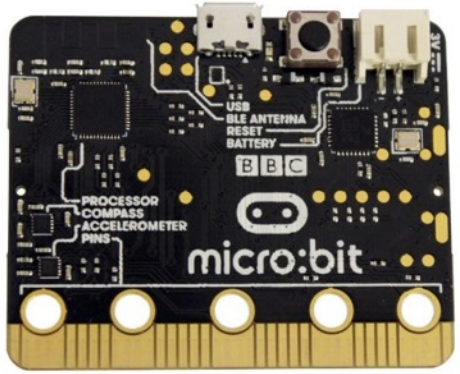
What is it?

- ”Swimming in sensors, drowning in data”
- Model design and optimization for highly constrained hardware platforms
- Can we get 10X+ reduction in ops or memory with minimal accuracy loss?

Near term results

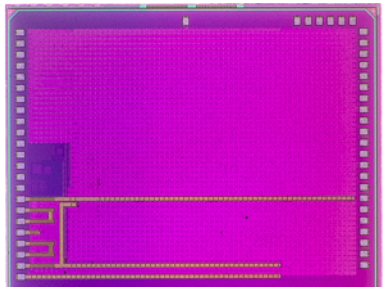
- Hybrid neural + non-neural techniques
- New training approaches for binary/ternary networks
- Compression techniques for recurrent neural networks (RNNs) that operate on time-series data

BBC Micro:Bit (Arm Cortex M0, 16KB RAM)



LPCXpresso 1125 (Arm Cortex M0, 8KB SRAM)

M0N0 (Arm Cortex M33, 16KB SRAM)



TinyML: HybridNet

"DS-CNN" is a highly optimized network for the key word spotting (KWS) task

- How do we optimize it further at iso-accuracy?

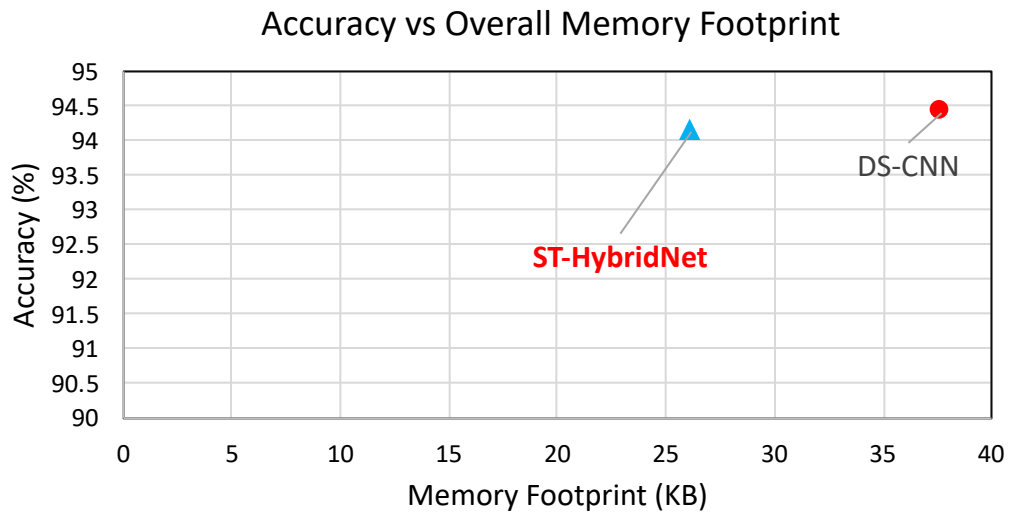
Ternarize weight values using Strassen's algorithm

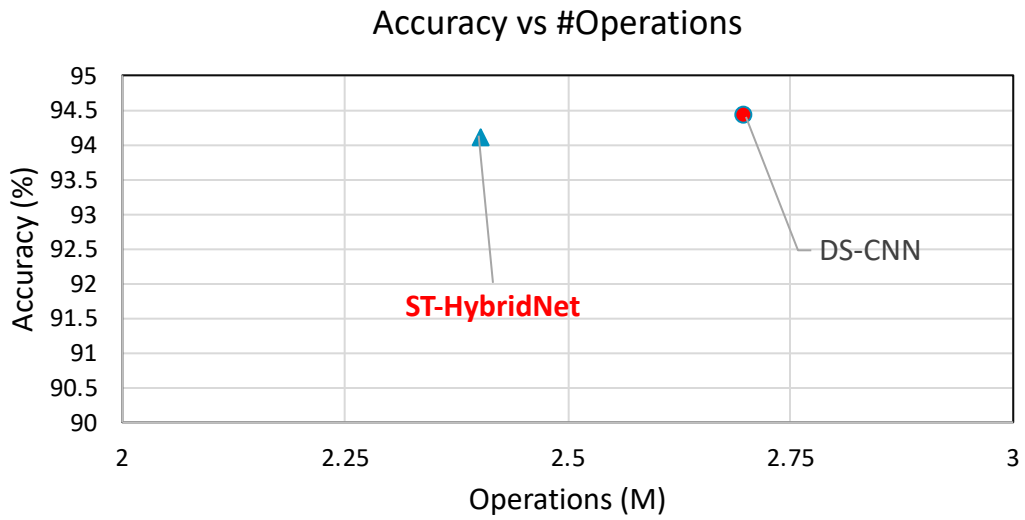
- Overall memory footprint reduced by 30%

Selectively use decision trees to reduce compute

- Total number of operations reduced by 12%

Less than 0.3% loss in accuracy for these savings





Published in SysML'19 - https://arxiv.org/abs/1903.01531

AutoBot

What is it?

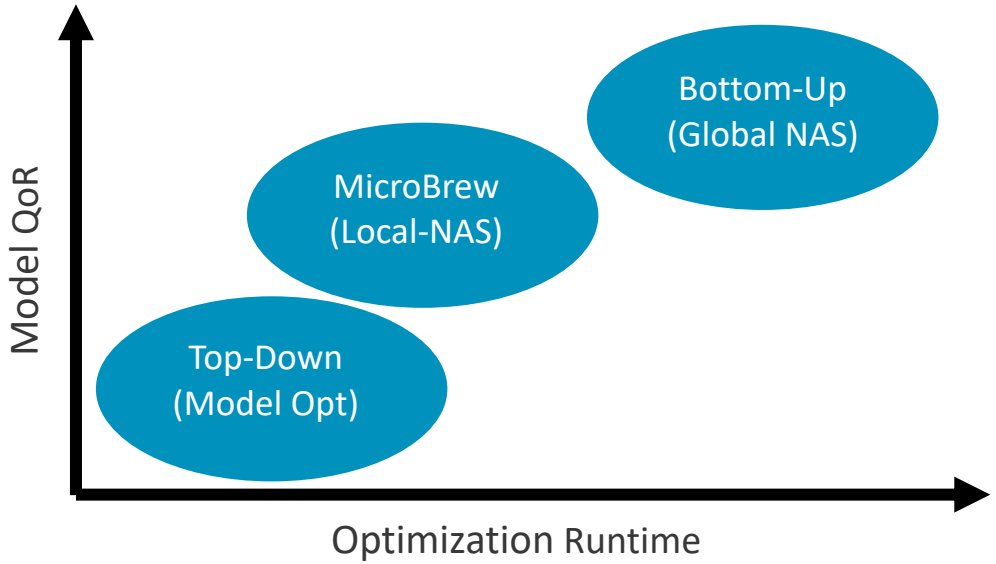
- Automate Neural Architecture Search (NAS) on Arm
- Incorporate information about Arm hardware into the optimization flow
- Reduce search runtime

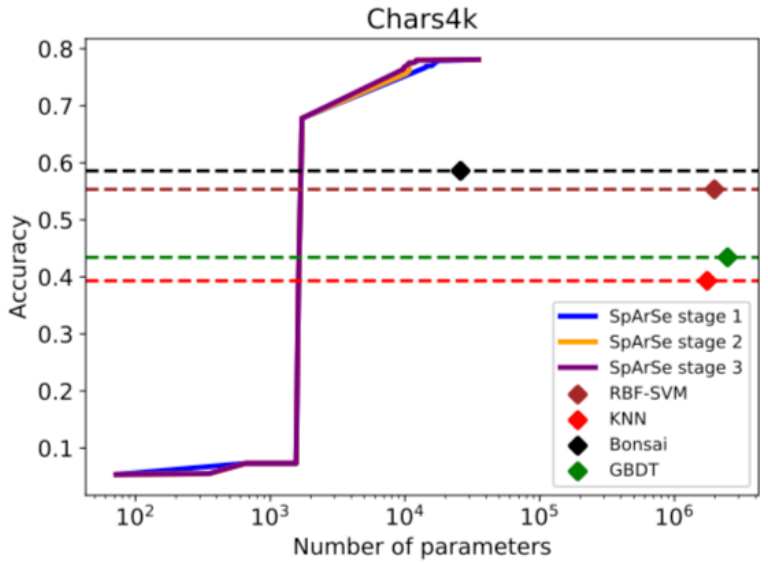
Near term goal: Top-Down (Optimization)

1. Input a trained model
2. Optimize for Arm IP – reduce latency/energy at iso-accuracy

Long term goal: Bottom-Up (Design)

1. Input a dataset
2. Create a from-scratch model optimized for Arm IP





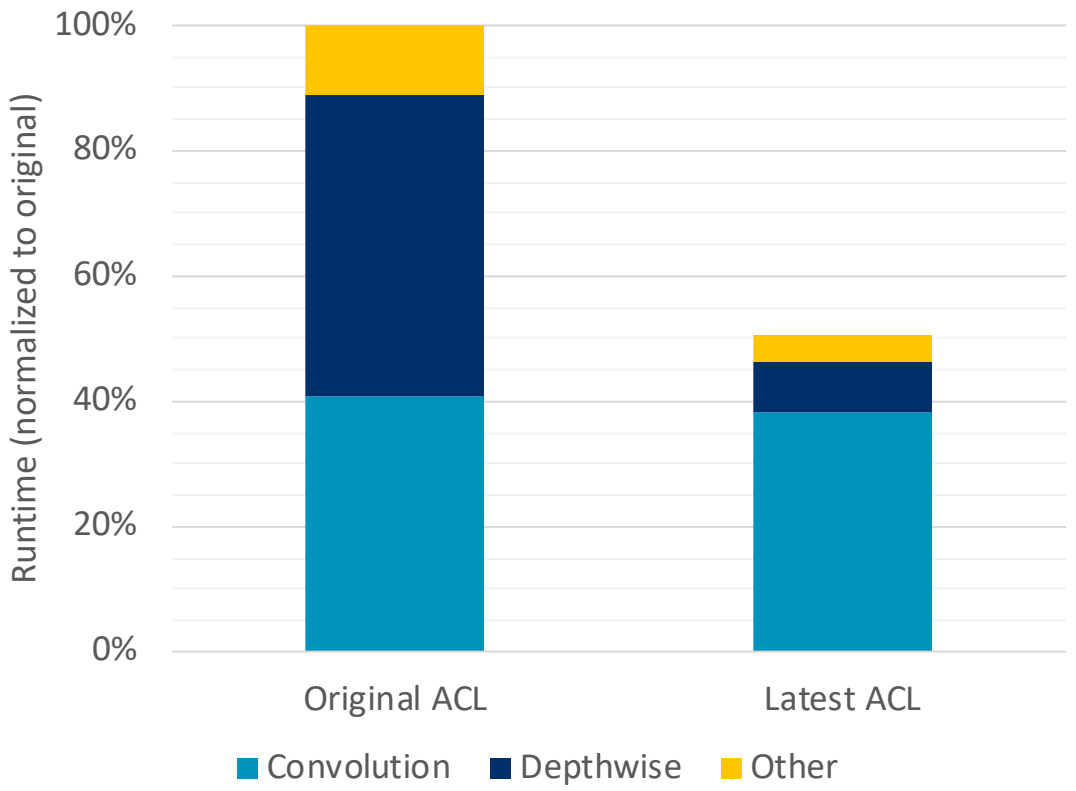
Accepted at NeurIPS'19 - https://arxiv.org/abs/1905.12107

ML Convolution Kernels in ArmCL

New optimized FP32 depthwise kernel

- Depthwise convolutions consist of depth-wise and point-wise layers
- RSH contributed new techniques for performing depthwise convolution
  - NEON optimized, cache-friendly, direct convolution outperforms GEMV based method
  - Activation fusion further reduces mem traffic
- 6x speedup compared to previous ArmCL FP32 depthwise kernels
- Overall 2x performance uplift for the whole MobileNet v1 model

Mobilenet v1 1.0/224 FP32; 1xA73



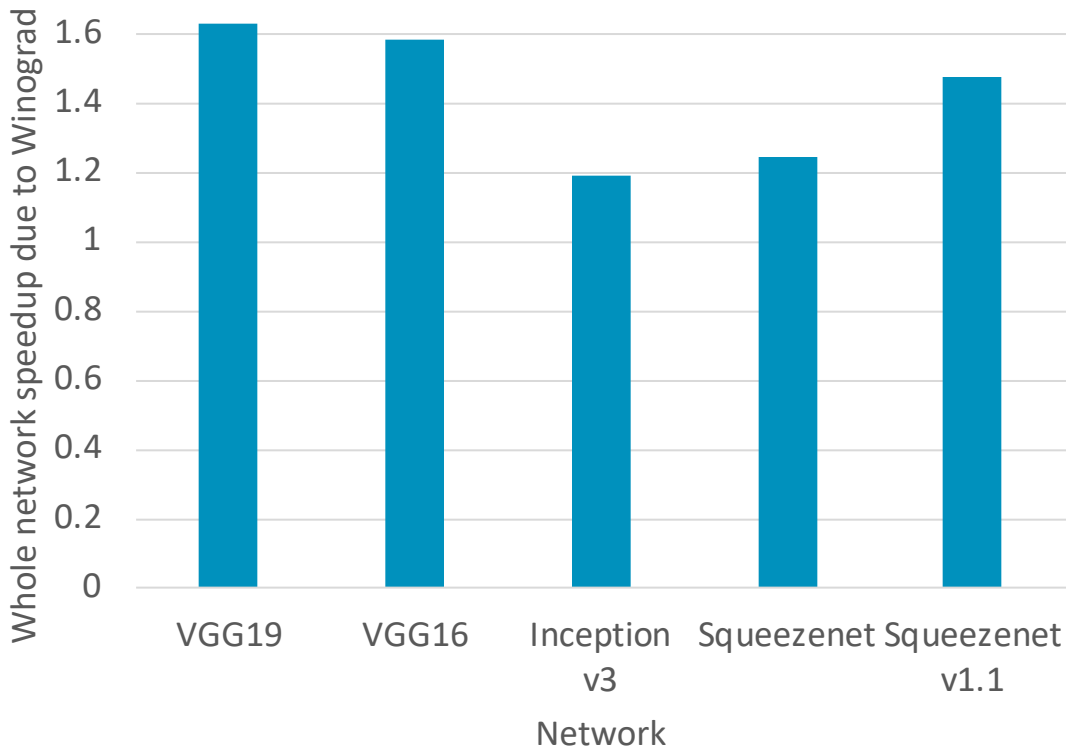
arm

ML Winograd Kernels in ArmCL

4x speedup to ArmCL FP32 convolution

- Introduced Winograd convolution to improve ArmCL CPU performance
- Winograd lowers matrix multiplication to element-wise multiplication
- Contributed code and analysis to ArmCL
- Up-to 1.6x whole network speedup
  - Depending on proportion of network time spent in convolution

ArmCL performance



arm

FixyNN

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

FixyNN: https://arxiv.org/abs/1902.11128

What is it?

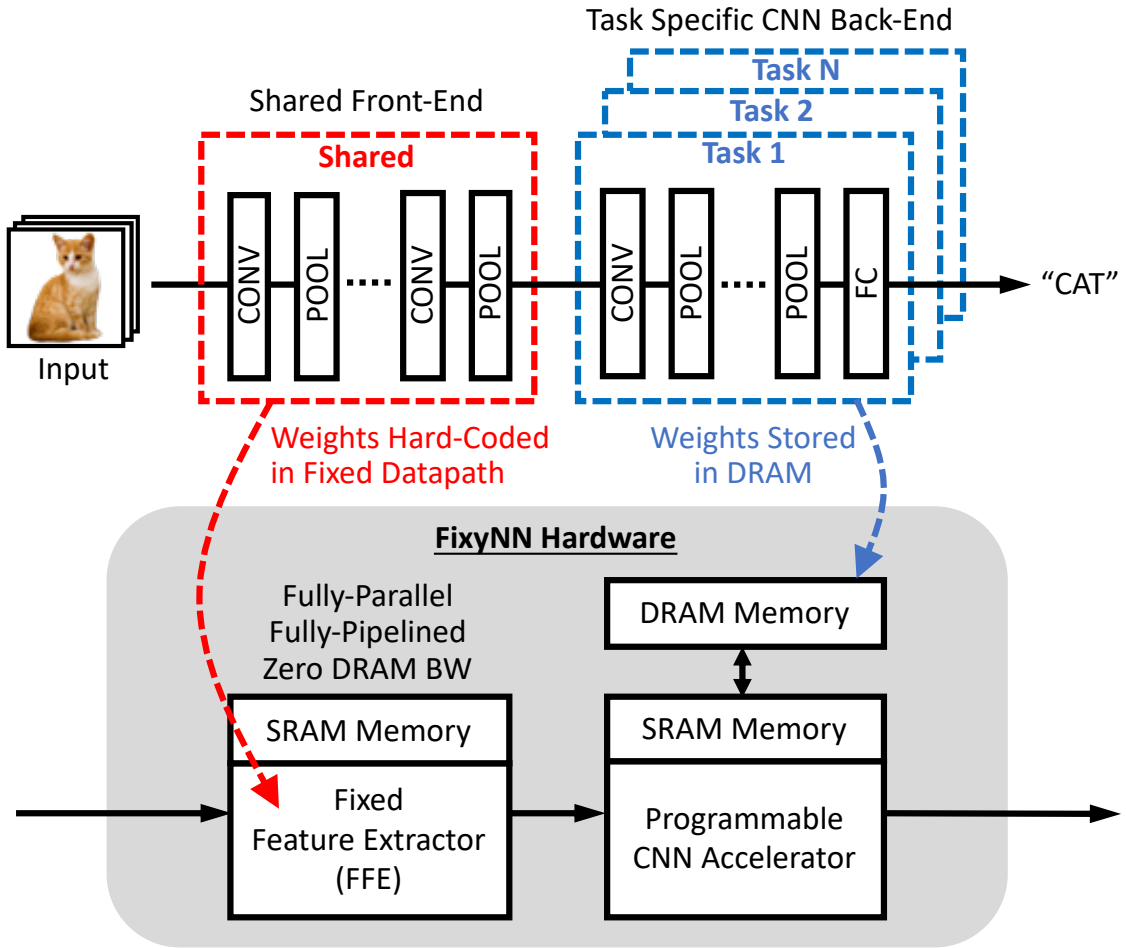
- Accelerator concept to push TOPS/W/mm^2

Goals

1. Aggressive HW specialization via transfer learning
2. Implement TF -> Verilog tool (DeepFreeze) to understand PPA benefit of fixed-weight datapaths
3. ML experiments to understand transfer learning
4. System model for iso-area comparison with baseline

Results

- Energy efficiency of up to 11.2 TOPS/W – nearly 2× more efficient than NVDLA alone in same area
  - **1.42x** TOPS/W by fixing 4/13 layers
  - **1.92x** TOPS/W by fixing 7/13 layers
- Accuracy loss of < 1% over six datasets



arm

arm

Discussion

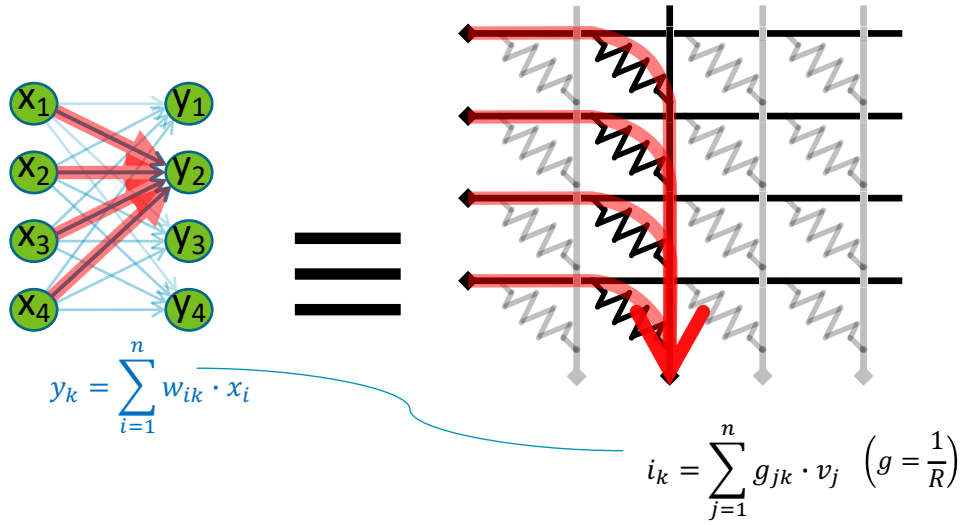
Analog HW and Non-neural models

The case for non-digital neural network accelerators

What are the most promising alternatives to digital CMOS?

What improvements in performance/W are possible?

What are the challenges? ADC/DACs? Noise?



arm

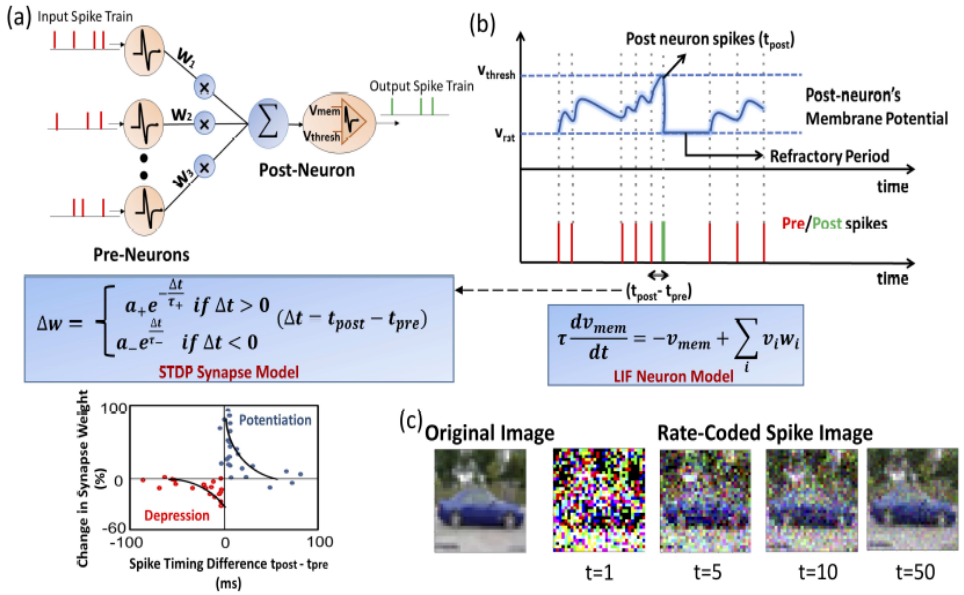
The case for non-neural network model architectures

What are the most promising alternatives to non-neural network models?

What problems do they solve?

In what ways are they better than neural models?

What are the challenges with non-neural network models?



From "Neuromorphic Computing: A bio-plausible route Toward spike-based machine intelligence", K. Ray et al

arm

Thank You!
Danke!
Merci!
谢谢!
ありがとう!
Gracias!
Kiitos!