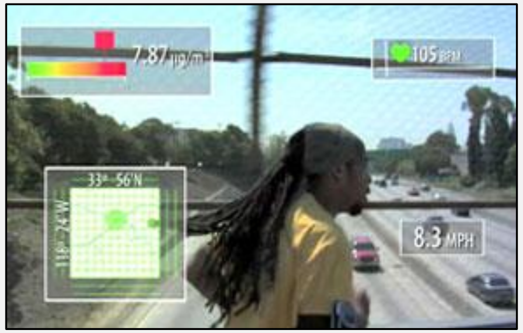
Samsung AI Forum
Seoul, South Korea
Tuesday 5th November 2019

The Deep *(Learning)* Transformation of Mobile and Embedded Computing

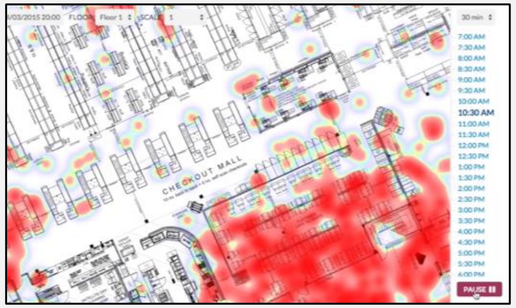
Nicholas D. Lane

@niclane7
http://mlsys.cs.ox.ac.uk



Mobile Health

Digital Assistants

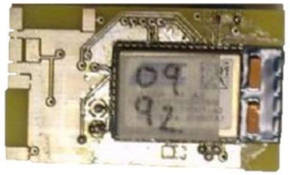


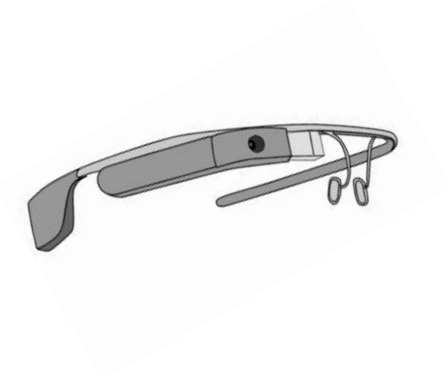
Quantified Enterprise

Urban Sensing

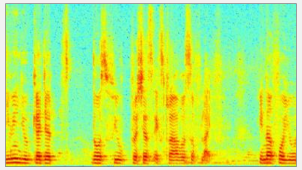
Consumer Personal Sensing

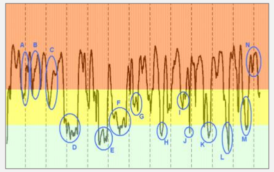
Sensor-driven Cities, Enterprises & Organizations





Audio Data





Inertial Data

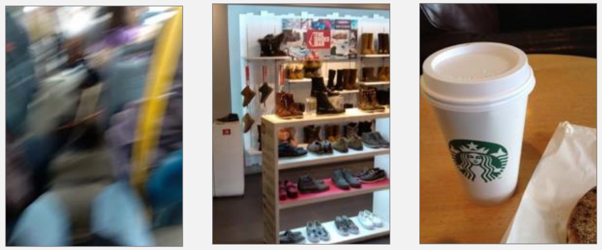
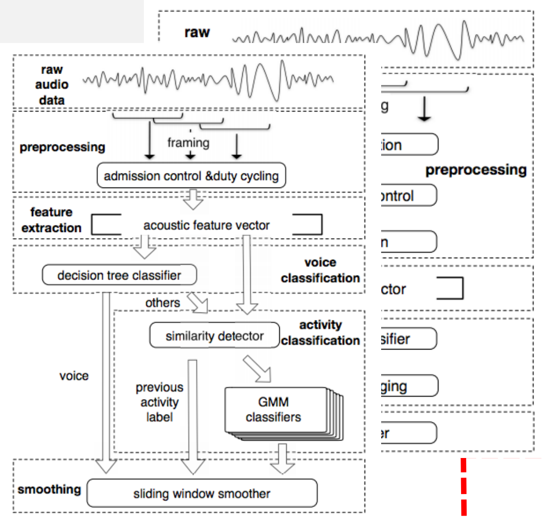


Image Data

Sensor Inference Pipelines



{stressed, not stressed}

{walking, running, sitting}

{music, conversation, male voice}

{shoes, subway, coffee cup}

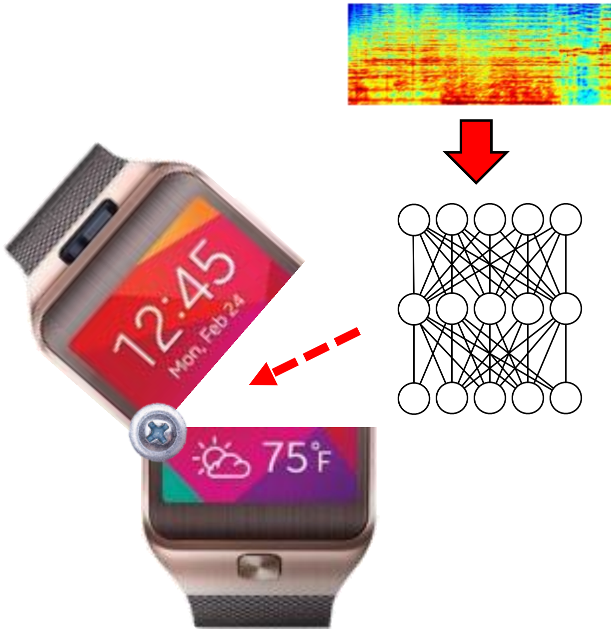
Sensors	Computation	Resources

Machine Learning is ***the*** core unifying building block that

spans all Mobile, Wearable, and Embedded Systems

Mobile and Embedded Deep Learning

AMBITION: Overcoming the system resource barriers that separate state-of-the-art ML and constrained classes of computing



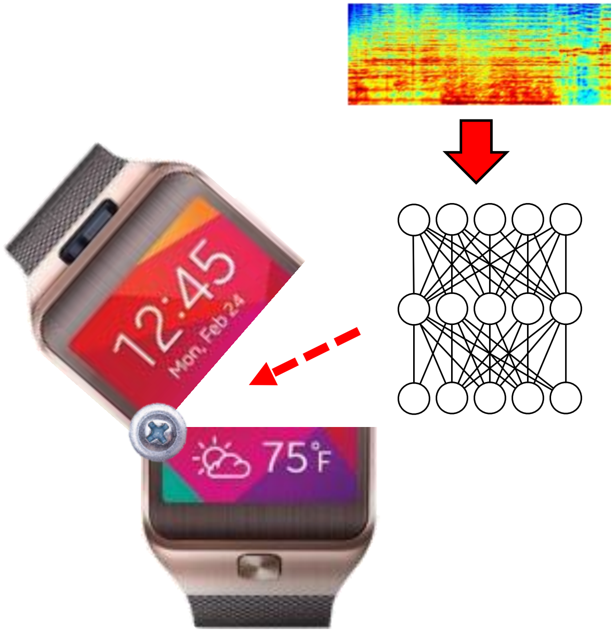
Next Frontier of Machine Learning

(1) Accuracy/Robustness

(2) Run Anywhere on Anything

Mobile and Embedded Deep Learning

AMBITION: Overcoming the system resource barriers that separate state-of-the-art ML and constrained classes of computing



Next Frontier of Machine Learning

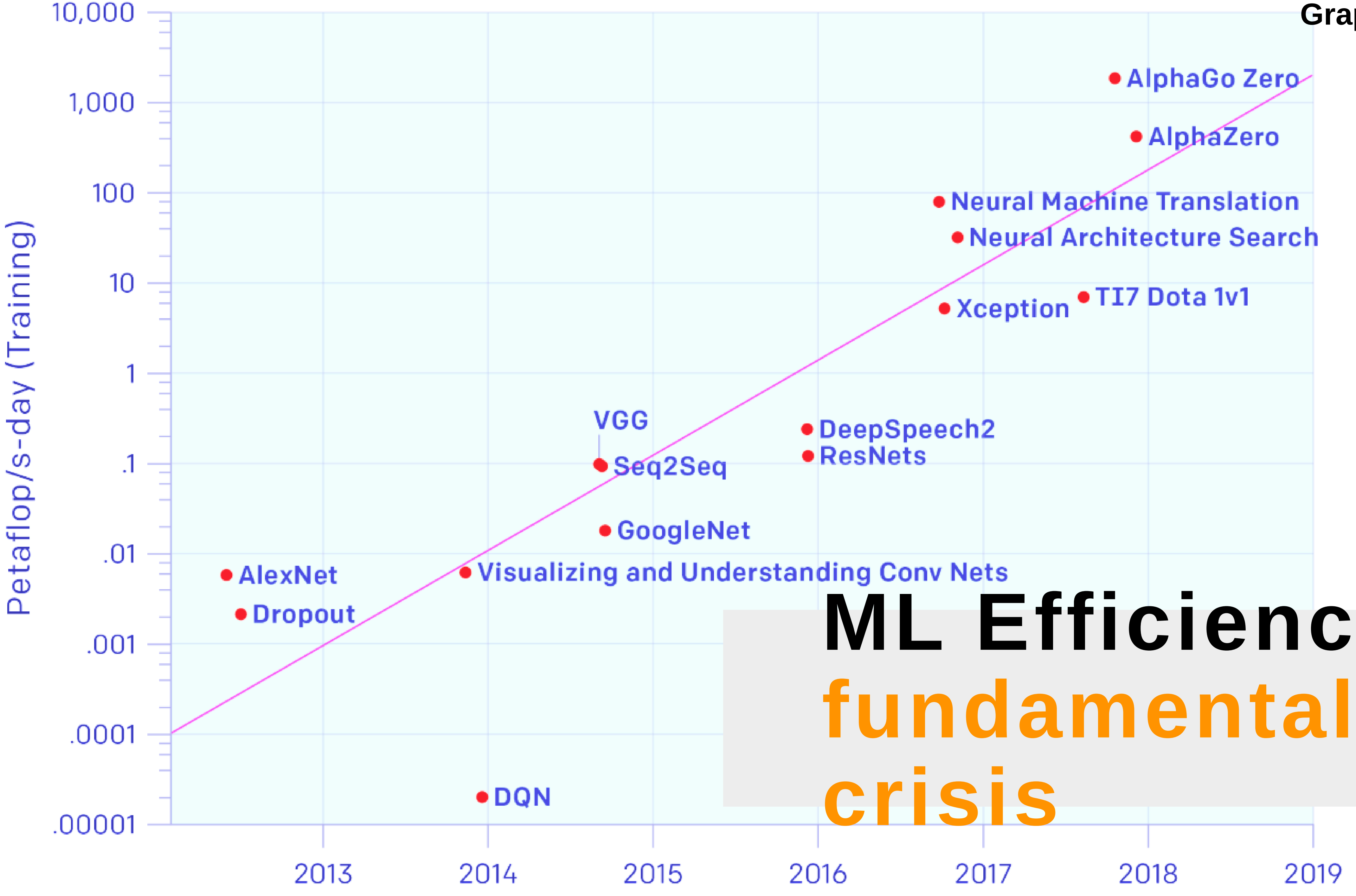
(1) Accuracy/Robustness

(2) Run Anywhere on Anything

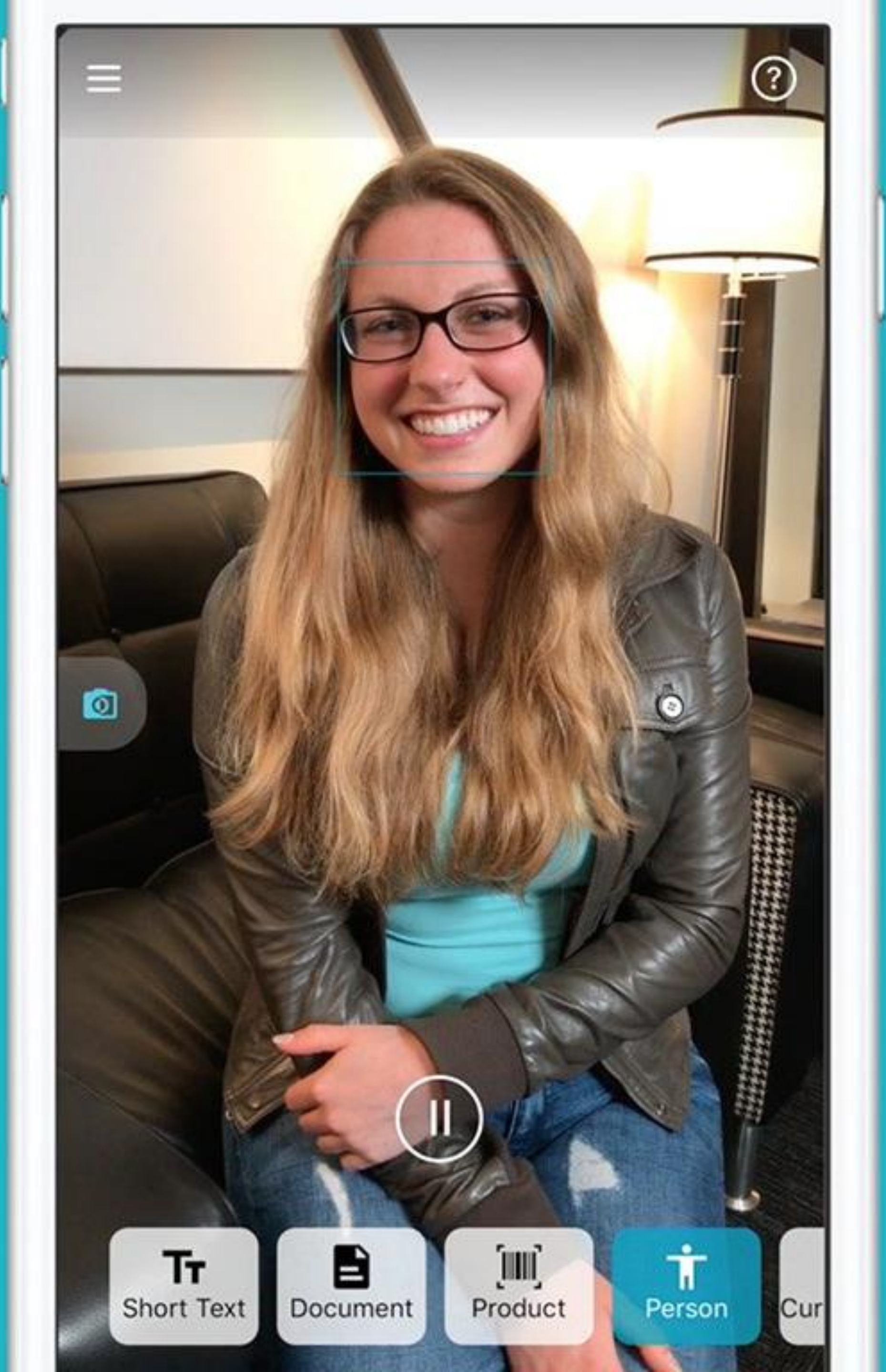
ML Efficiency drives device capabilities

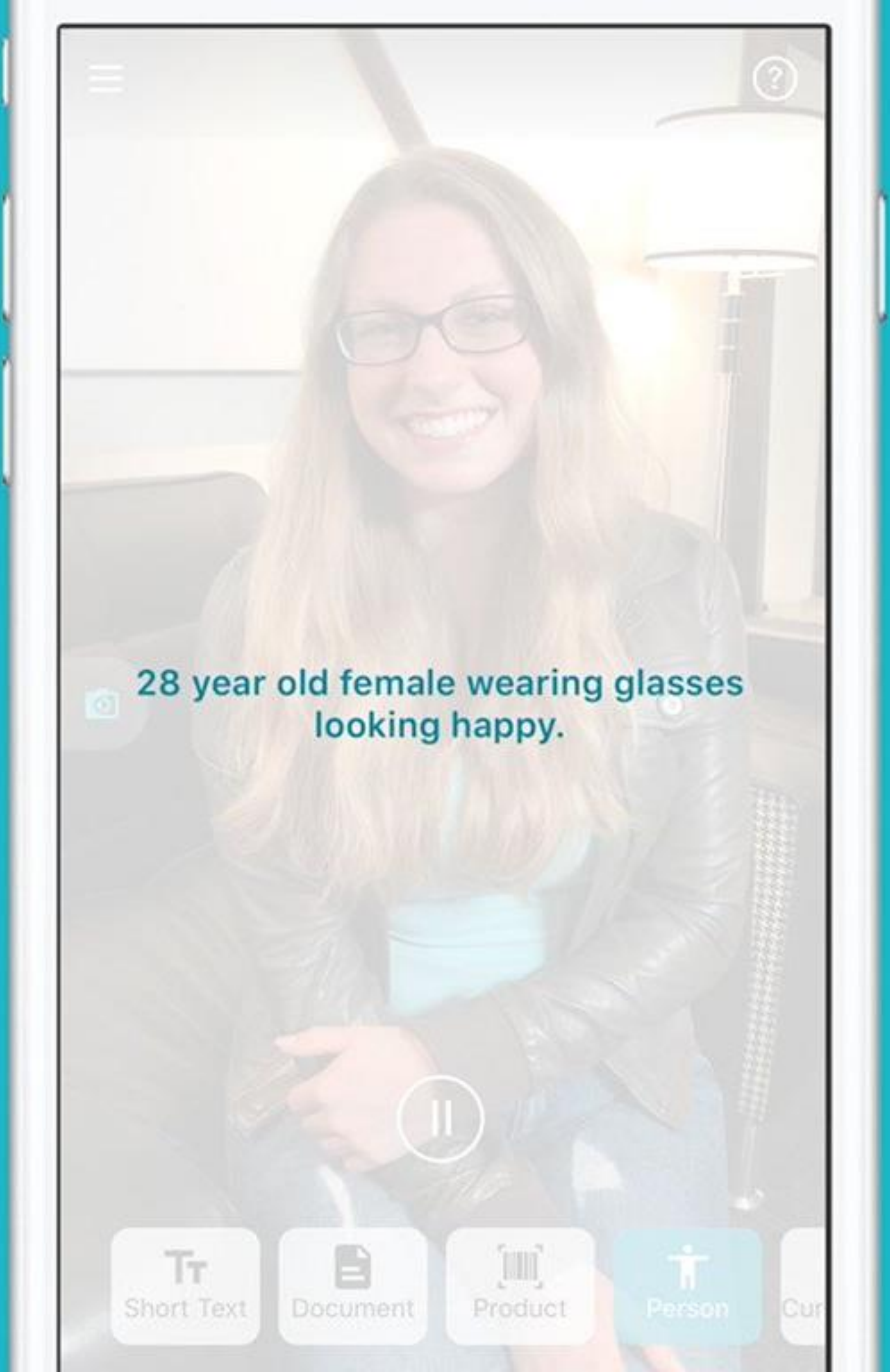
- Enabling state-of-the-art techniques across all systems
- **USER PRIVACY**
- No need for developing a range of simple and complex ML models
- Real-time Execution *(without dependency on network connectivity)*
- Model Size *(think: updating a mobile app if model alone is 500MB)*

Graphic Ack. OpenAI



ML Efficiency is a fundamental crisis

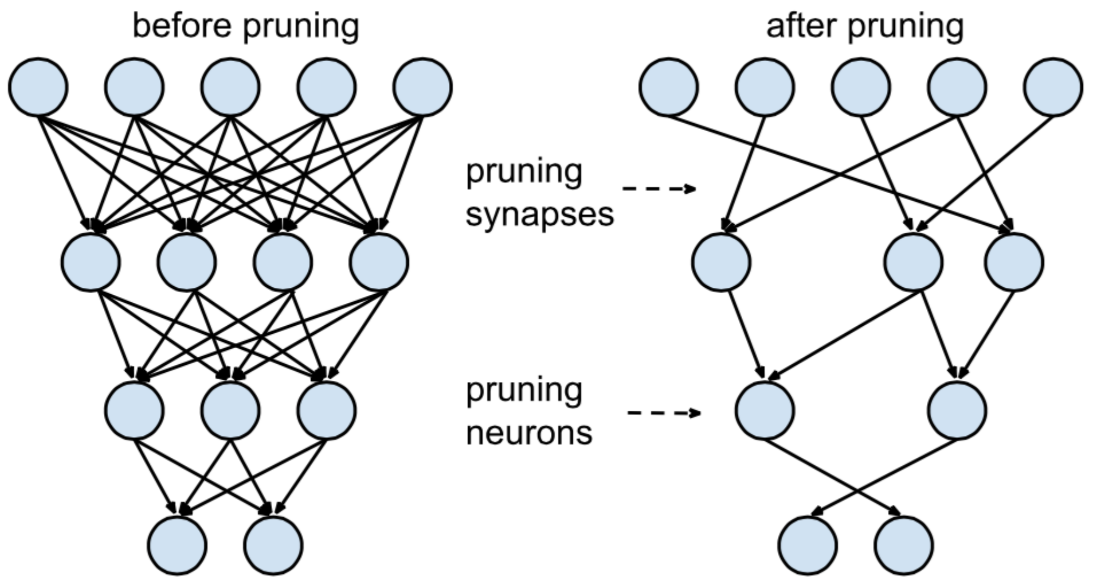

Short Text
Document
Product
Person

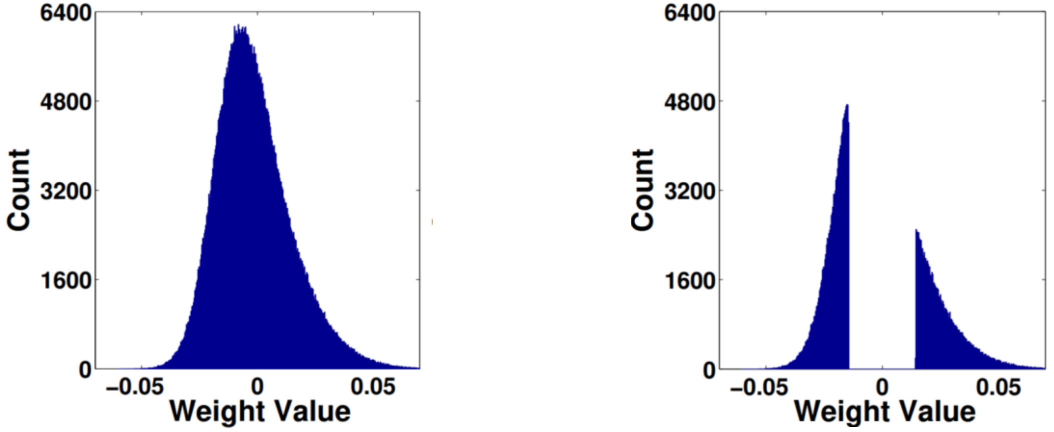

28 year old female wearing glasses looking happy.
Short Text
Document
Product
Person

Node Pruning

Many heuristics developed to determine which nodes to prune

Example: Prune nodes with absolute weights below a threshold

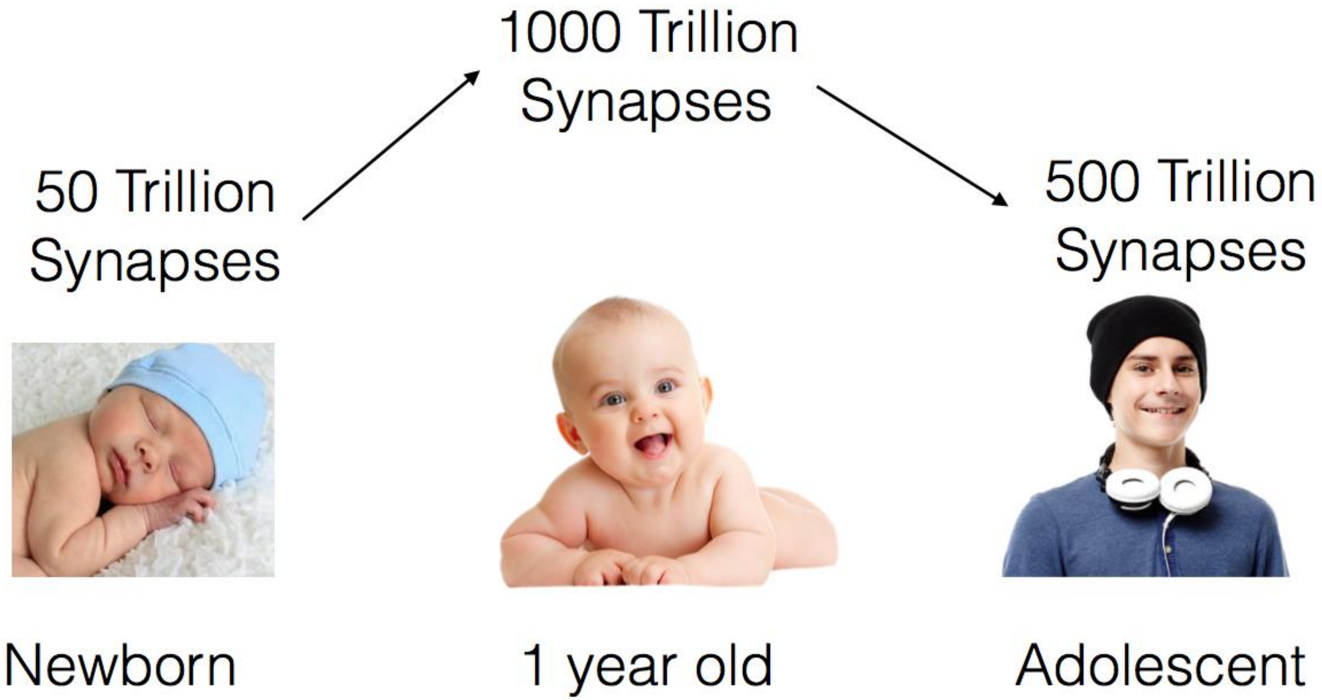




Song Han, Jeff Pool, John Tran, William J. Dally, “Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015

Grounding in Nature?

Number of synapses in the human brain during child development



Christopher A Walsh, Peter Huttenlocher (1931 – 2013). Nature, 502(7470):172-172, 2013.

Slide ACK Song Han

Starting in Late 2014: Mobile & Embedded DL

2014

1st Proof-of-Concept DL on Mobile **[HotMobile ‘15]**

2015

DeepEar (1st DSP-based DL General Audio Sensing) **[UbiComp ‘15]** *Best Paper*

DL Smartwatch Activity Recognition **[WristSense ‘16]** *Best Paper*

2016

1st time: VGG executing directly on a commodity SmartWatch

1st time: Smartphone-scale DL on embedded processors (e.g., M0/M3) **[SenSys ‘16]**

2017

1st time: Multiple DL Vision Models on Wearable **[MobiSys ‘17]**

Notable Additional Innovations

Algorithmic & Architecture Advances

- Node Pruning
- SqueezeNet (50x AlexNet reduction)
- Low Precision Results (8-bit etc)
- Binarization of Networks
- MobileNet, Small-footprint Nets

Hardware Innovations

- Diannao and Cnvlutin2
- Front-ends e.g., SNPE - Qualcomm
- TPU, FPGAs / Hybrids
- Analog from Digital Approaches
- Spiking H/W & Approx. Compute

Starting in Late 2014: Mobile & Embedded DL

2014

1st Proof-of-Concept DL on Mobile **[HotMobile '15]**

2015

DeepEar (1st DSP-based DL General Audio Sensing) **[UbiComp '15]** *Best Paper*

DL Smartwatch Activity Recognition **[WristSense '16]** *Best Paper*

2016

1st time: VGG executing directly on a commodity SmartWatch

1st time: Smartphone-scale DL on embedded processors (e.g., M0/M3) [SenSys '16]

2017

1st time: Multiple DL Vision Models on Wearable **[MobiSys '17]**

Notable Additional Innovations

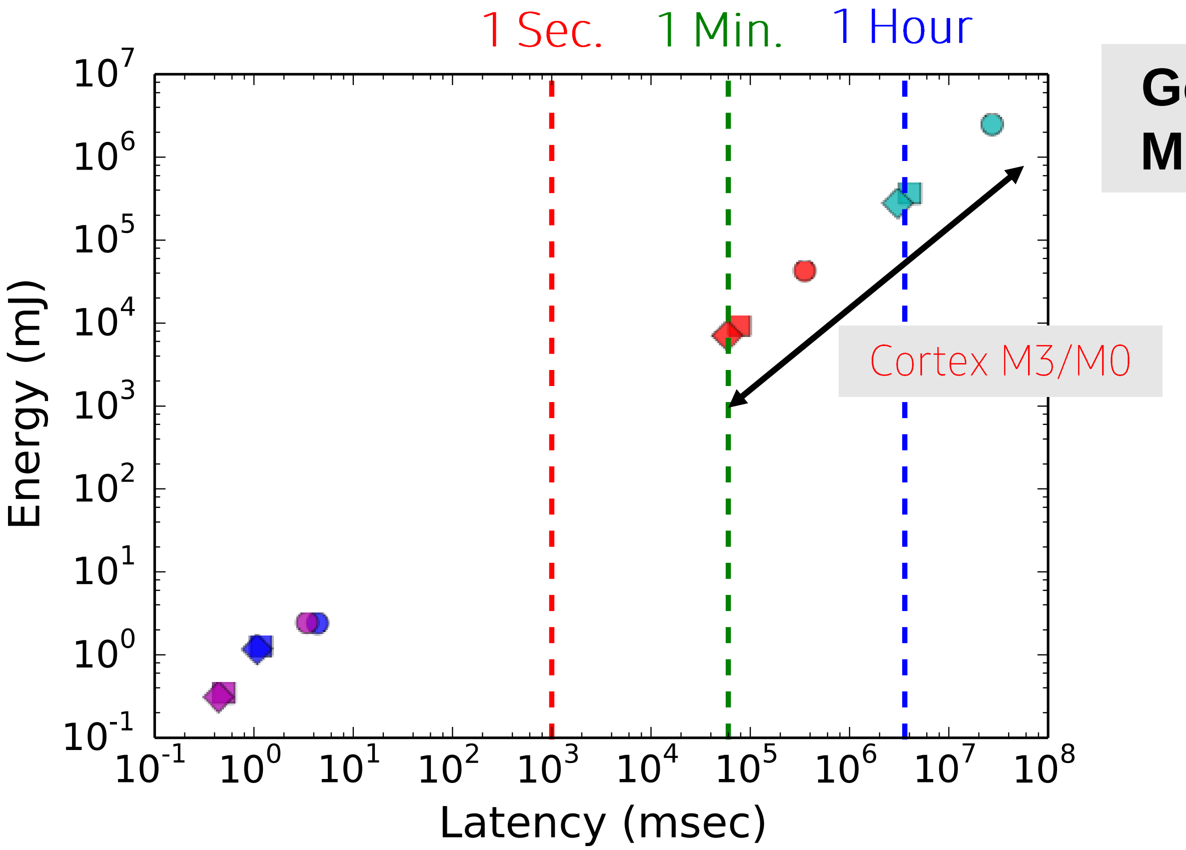
Algorithmic & Architecture Advances

- Node Pruning
- SqueezeNet (50x AlexNet reduction)
- Low Precision Results (8-bit etc)
- Binarization of Networks
- MobileNet, Small-footprint Nets

Hardware Innovations

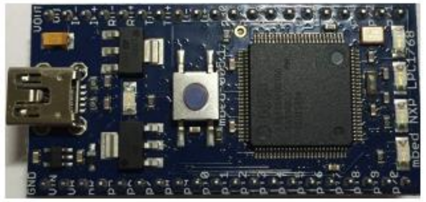
- Diannao and Cnvlutin2
- Front-ends e.g., SNPE - Qualcomm
- TPU, FPGAs / Hybrids
- Analog from Digital Approaches
- Spiking H/W & Approx. Compute

Early 2016: Deep Learning on Microcontrollers



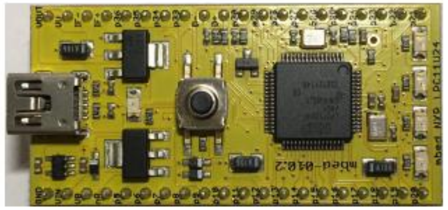
Google SpeakerID Model (FC Layers)

32 KB



ARM Cortex M3

16 KB



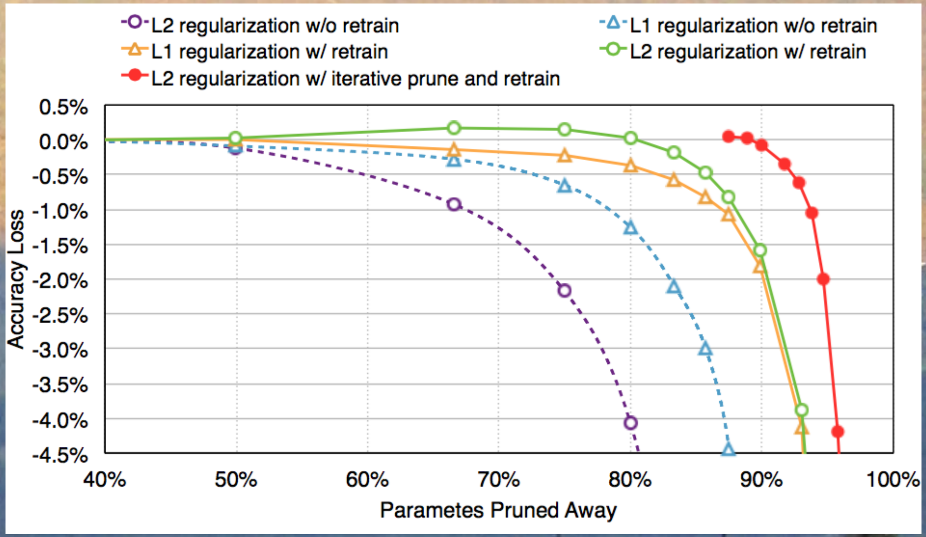
ARM Cortex M0

2-4% degradation in accuracy

Sourav Bhattacharya, Nicholas Lane, “Sparsifying Deep Learning Layers for Constrained Resource Inference on Wearables", SenSys 2016


9N-AHA

Obsession with Model Compression



The first 50x gains were "easy."

But where will I find my next 50x?

Algorithmic & Architecture Advances

- Node Pruning
- SqueezeNet (50x AlexNet reduction)
- Binarization, Low Precision (8-bit etc)
- MobileNet, Small-footprint Nets

Hardware-centric Innovations

- Diannao and Cnvlutin2
- Front-ends e.g., SNPE - Qualcomm

The first 50x gains were “easy.”

But where will I find my next 50x?

Forgotten 1st-Gen Methods

Algorithmic & Architecture Advances

- Node Pruning
- SqueezeNet (50x AlexNet reduction)
- Binarization, Low Precision (8-bit etc)
- MobileNet, Small-footprint Nets

Hardware-centric

- Front-ends e.g., SNPE - Qualcomm

The first 50x gains were "easy."

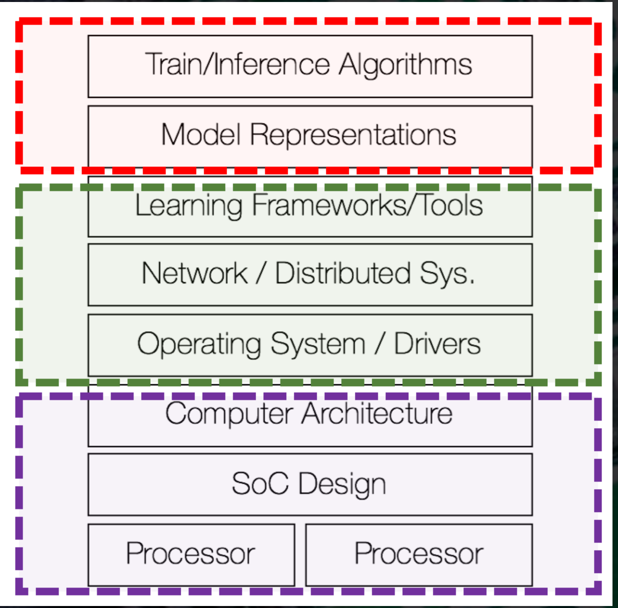
But where will I find my next 50x?

Fundamental On-Device ML Challenges

#1: Modular Low-data Movement Learning Algorithms

#2: Automated Specialization

#3: Memory and Compute Sharing



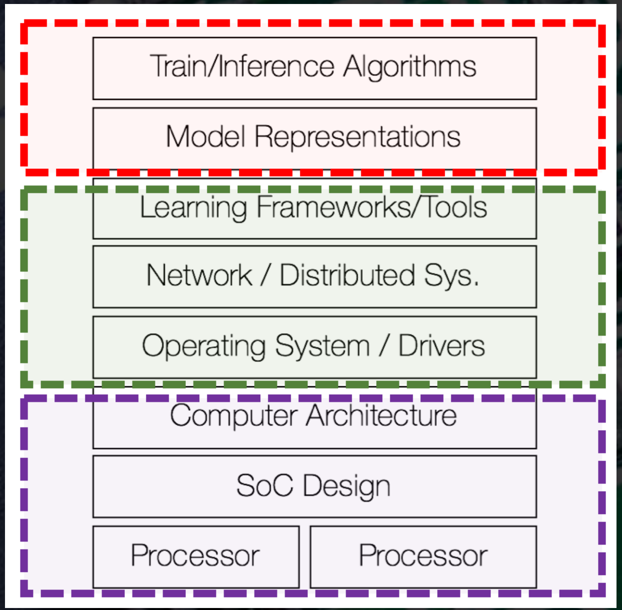
Rethinking the complete stack (and the learning algorithms)

Fundamental On-Device ML Challenges

#1: Modular Low-data Movement Learning Algorithms

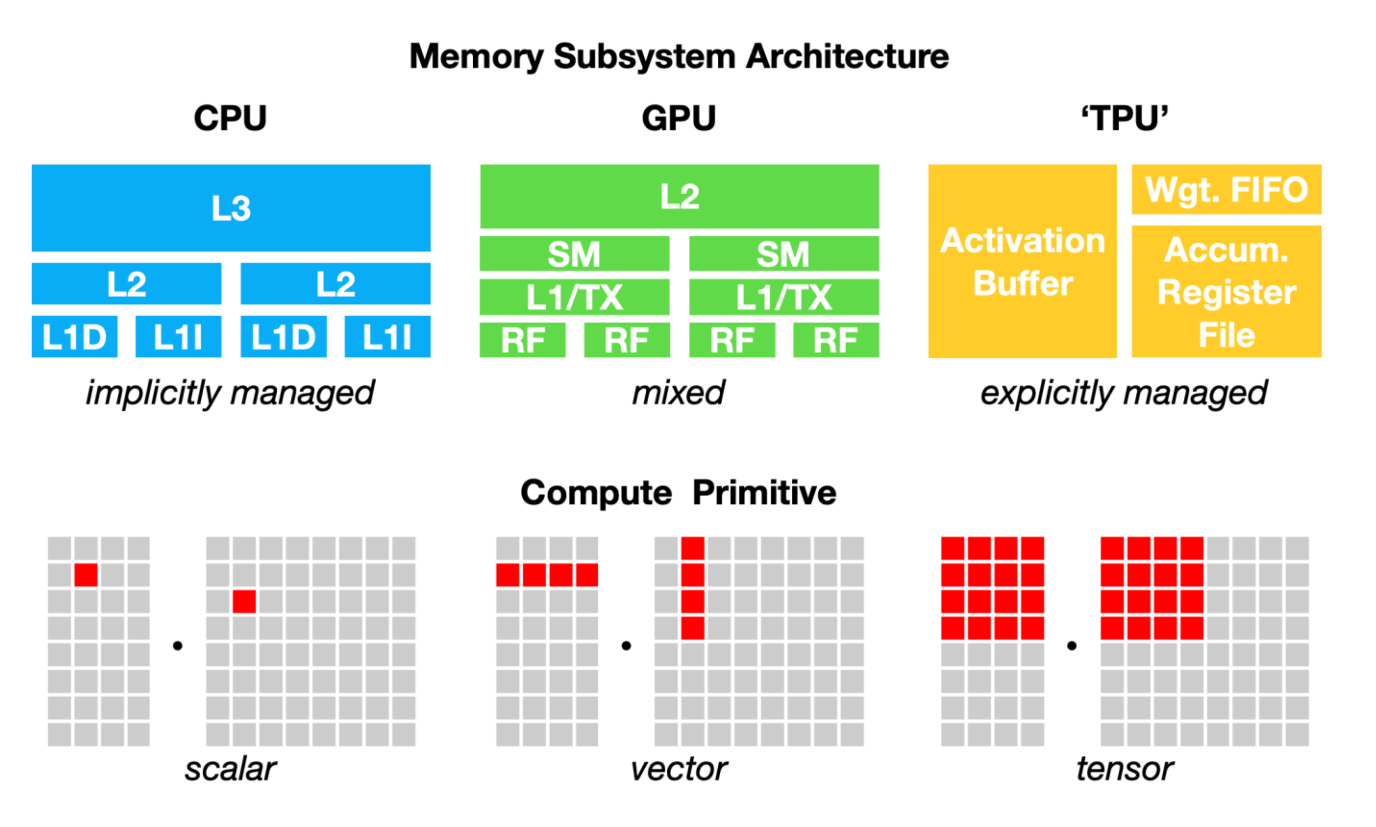
#2: Automated Specialization

#3: Memory and Compute Sharing

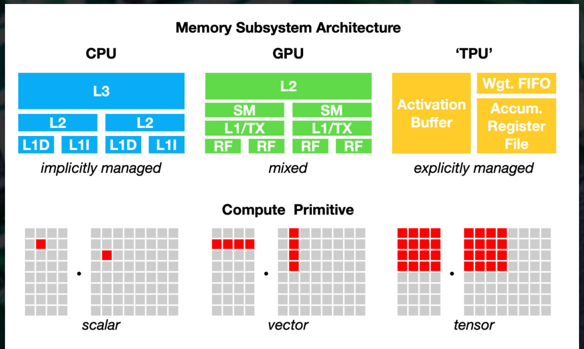


Rethinking the complete stack (and the learning algorithms)

#2 Automated Specialization

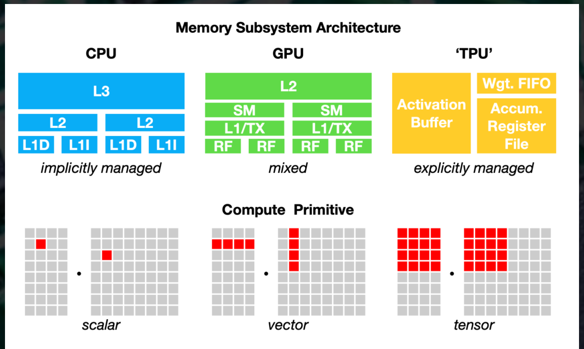


#2 Automated Specialization

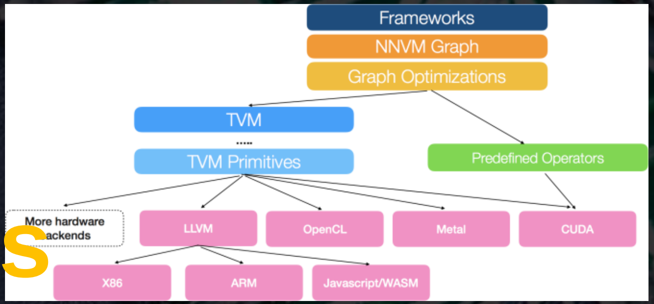


Vanilla AutoML output

#2 Automated Specialization

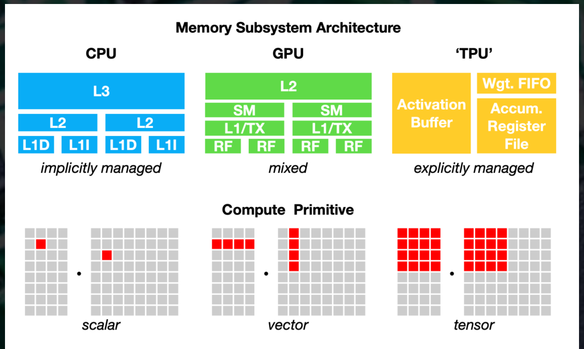


Vanilla AutoML output

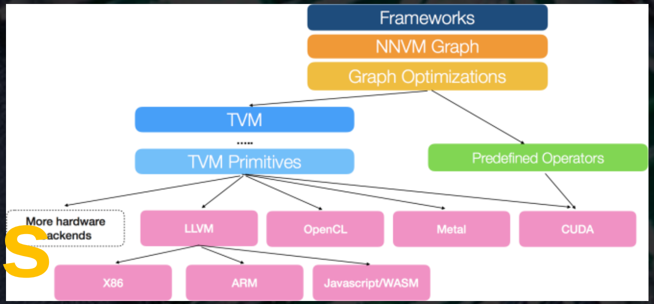


DL Compilers

#2 Automated Specialization

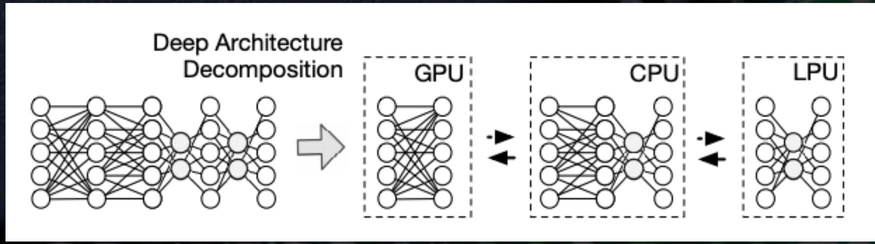


Vanilla AutoML output



DL Compilers

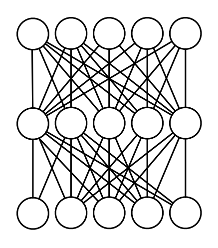
Semi Hand-built Examples



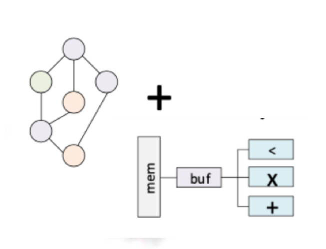
Nicholas Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, Lei Jiao, Lorena Qendro, Fahim Kawsar, "DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices", IPSN 2016

#2 Automated Specialization

- Combine: Optimization, AutoML and Code-Gen
- Model deeply SoC behavior, beyond constraint searching due to memory and FLOPS
- Automation allows for: per-model per-task per-device
- Integrate hooks and meta-data for runtime efficiency



Hardware Specialization



AMBITION: Automated offline generation of ML models specialized for a target chip/platform that rivals hand-design

Automated Specialization Example: Huge Drop in Audio Sensing Latency under **Automated Mobile GPU Tuning**

Audio Processing Pipelines

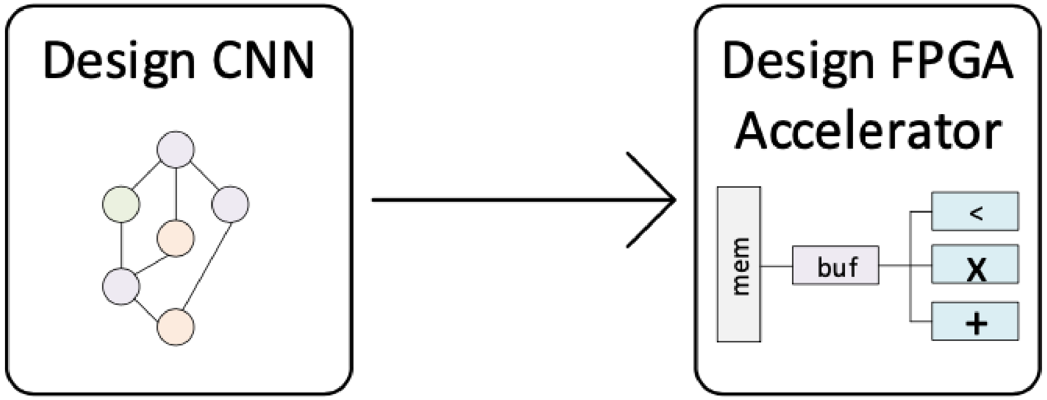
	GMM *[full pipeline]*	GMM *[model only]*	DNN *[full pipeline]*	DNN *[model only]*
DSP	-8.8x	-8.6x	-4.5x	-4.0x
DSP-*m*	-3.2x	-2.5x	-2.1x	-1.5x
CPU	1.0x (1573*ms*)	1.0x (1472*ms*)	1.0x (501*ms*)	1.0x (490*ms*)
CPU-*m*	3.0x	3.4x	2.8x	2.9x
***n*-GPU**	3.1x	3.6x	1.8x	1.8x
***a*-GPU**	**8.2x**	**16.2x**	**13.5x**	**21.3x**

Petko Georgiev, Nicholas Lane, Cecilia Mascolo, David Chu, "Accelerating Mobile Audio Sensing Algorithms through On-Chip GPU Offloading", MobiSys 2017

Platform

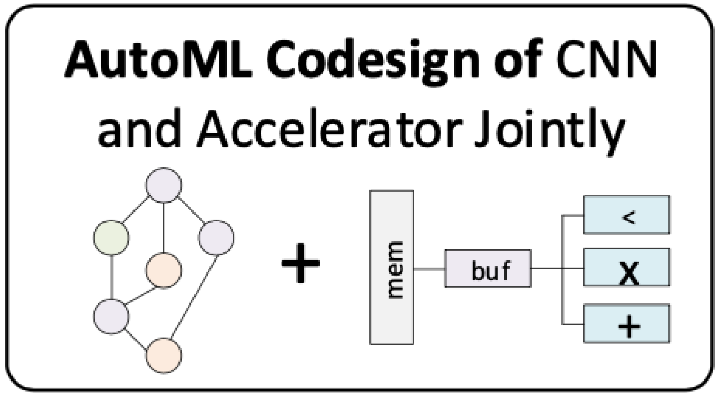
Qualcomm Snapdragon 800

Automated Specialization Example: **Joint Optimization** of Accelerator Design and Deep Neural Architecture



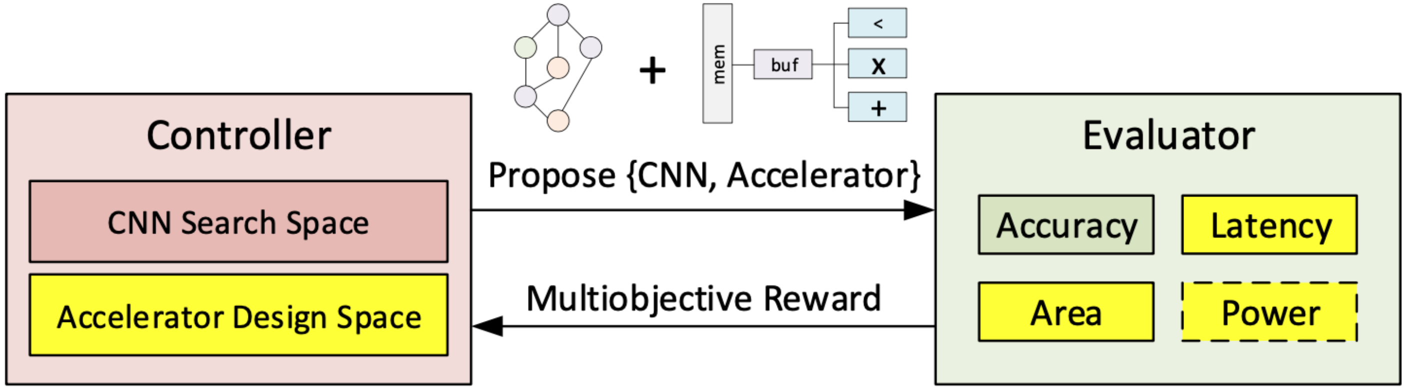
conventional approach

vs

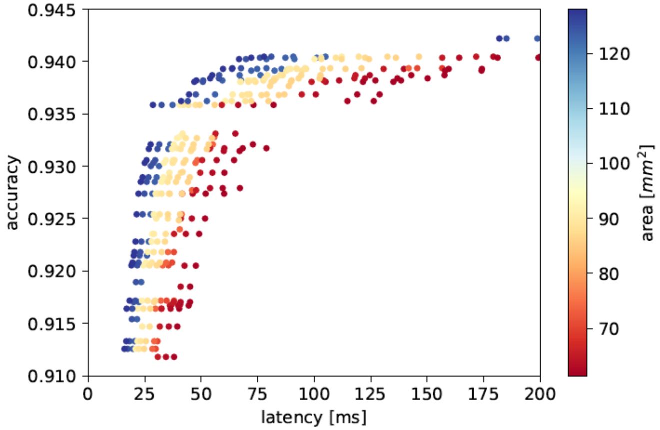


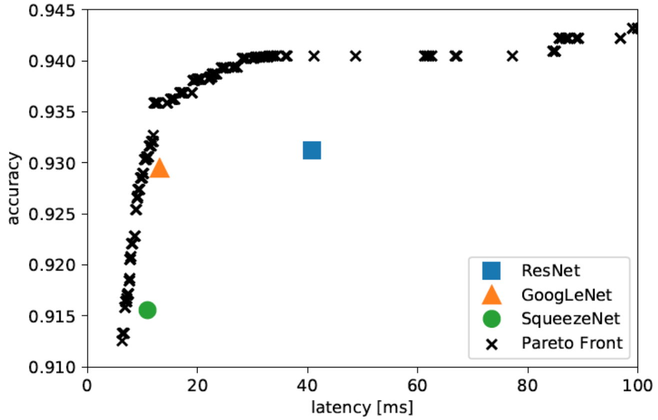
joint optimization

Automated Specialization Example: Joint Optimization of Accelerator Design and Deep Neural Architecture



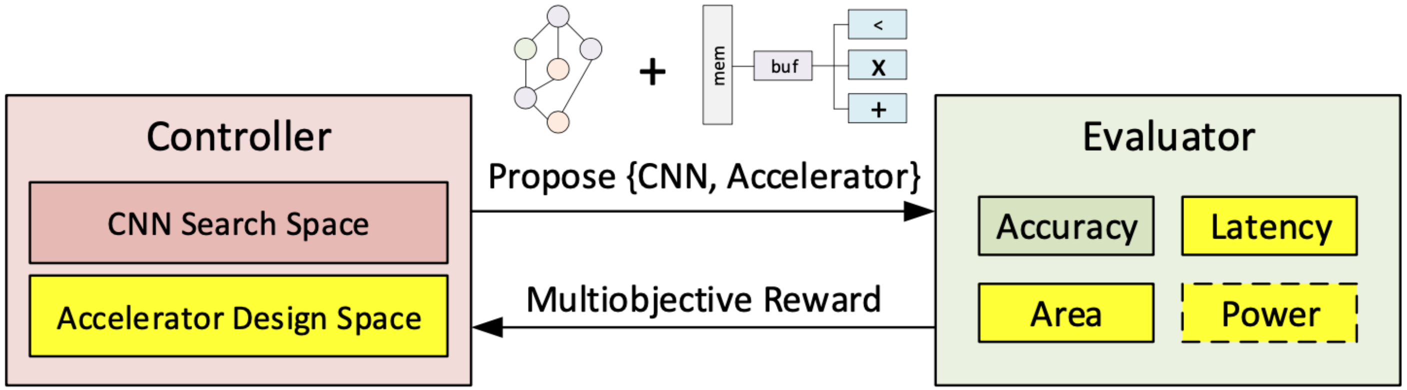
Platform Zync Ultrascale+





Mohamed Abdelfattah, Lukasz Dudziak, Thomas Chau, Hyeji Kim, Royson Lee, Nicholas D. Lane, "Best of Both Worlds: AutoML Codesign of a CNN and its FPGA Accelerator", *under submission ISFPGA '20*

Automated Specialization Example: Joint Optimization of Accelerator Design and Deep Neural Architecture

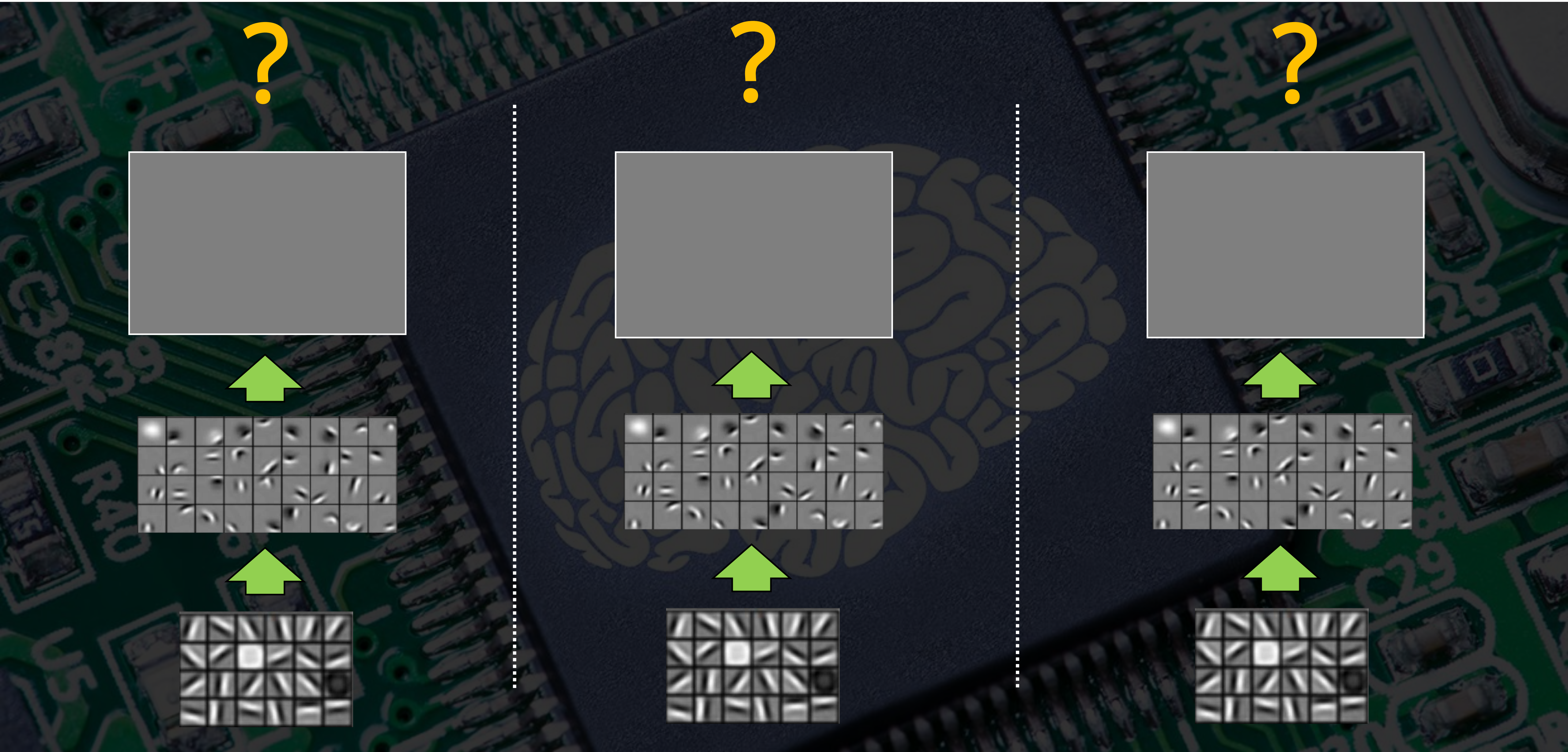


Platform Zync Ultrascale+

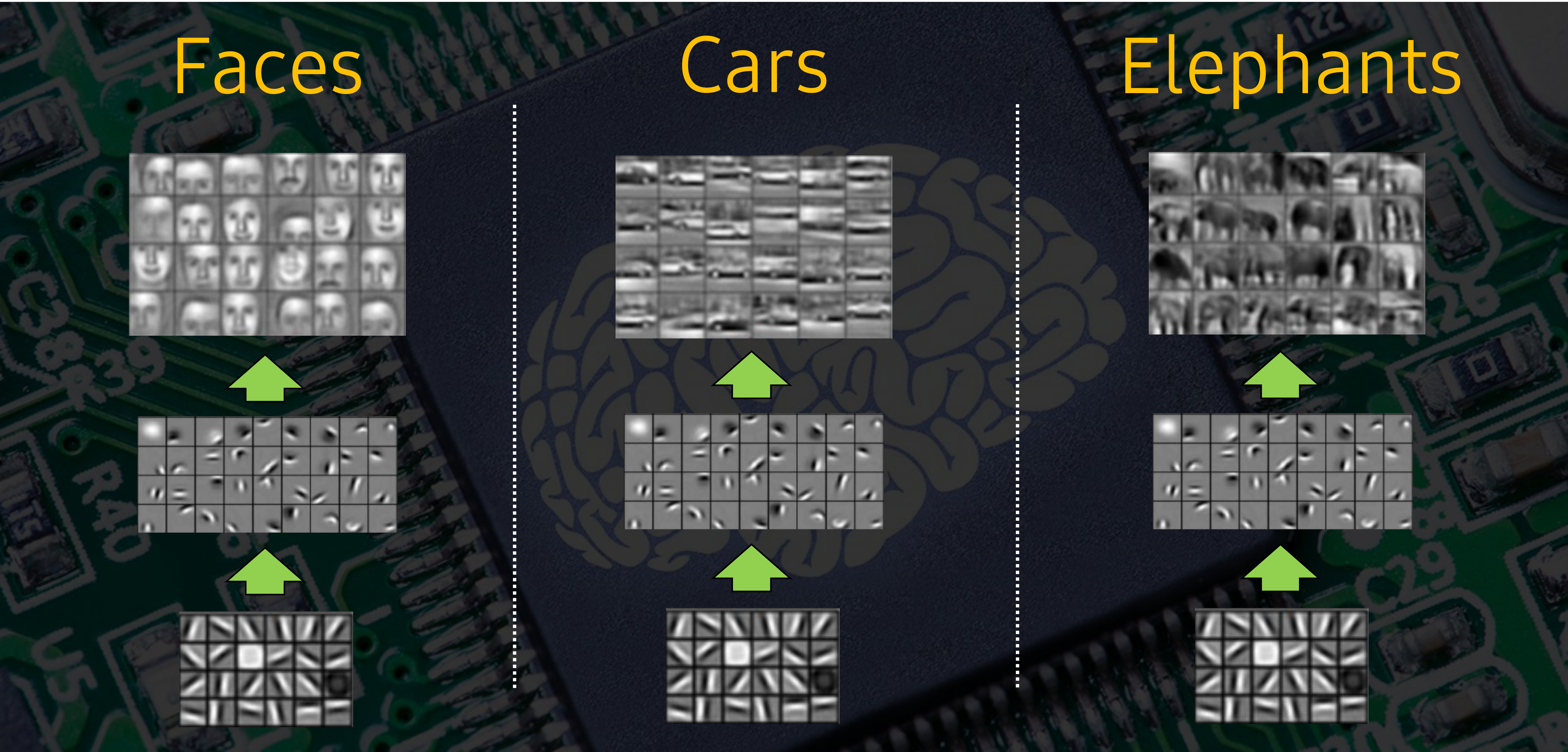
	prior SOA	HWNAS
Accuracy	92.8%	93.6%
Latency	51ms	42ms
HW Area	170	130

Mohamed Abdelfattah, Lukasz Dudziak, Thomas Chau, Hyeji Kim, Royson Lee, Nicholas D. Lane, "Best of Both Worlds: AutoML Codesign of a CNN and its FPGA Accelerator", *under submission ISFPGA '20*

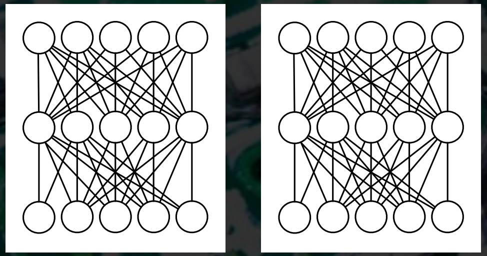
#3 Memory and Compute Sharing



#3 Memory and Compute Sharing



#3 Memory and Compute Sharing

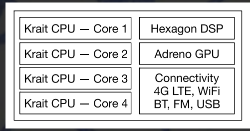


Trained Models

- **Schedulers**
- **Partitioned CPU/xPU Execution (including offloading)**
- **Memory Layout and Context Switching**
- **Micro-kernels for management of NPUs etc.**
- **Initialization of Accelerators and Hetero Compute**

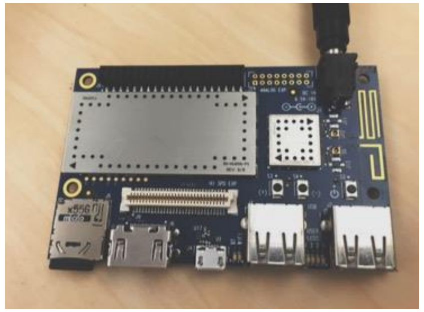
ML-aware Systems Components

Runtime Resources



AMBITION: Maximize runtime resource utilization through the ML-aware sharing and scheduling memory & compute

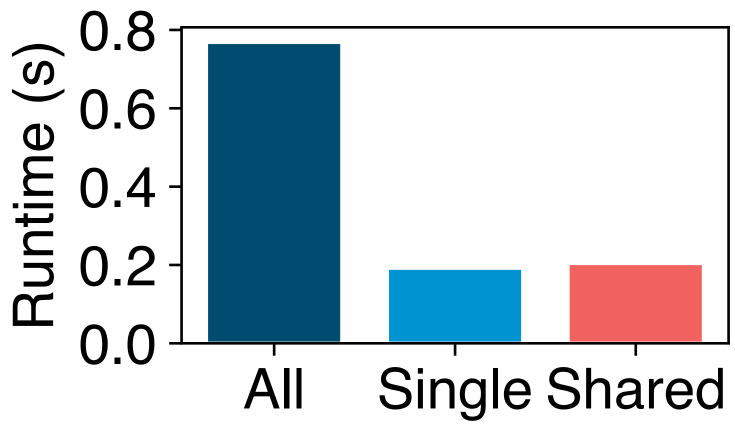
Sharing Resource Example: Scaling to **Multiple Audio Tasks** w/ Negligible Loss in **Accuracy**



Qualcomm
Snapdragon 400

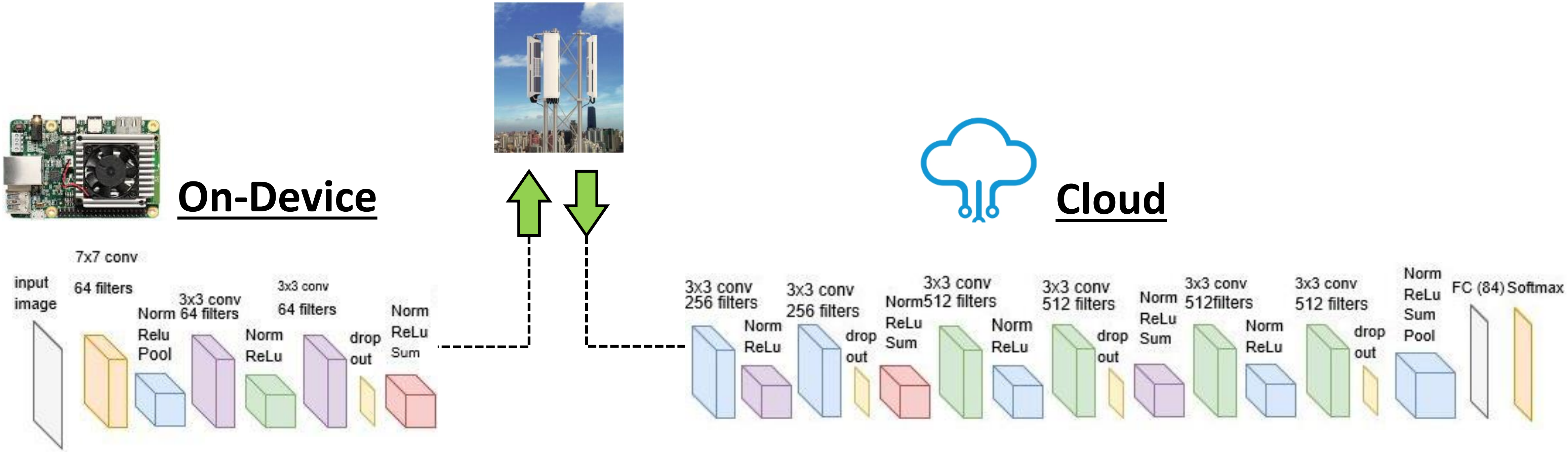
	Single Model	Avg. Multi-Task Model
Speaker Identification	85.1%	84.7 (±1.2%)
Emotion Recognition	83.4%	85.8 (±1.6%)
Stress Detection	85.4%	83.3 (±2.0%)
Ambient Scene Analysis	84.8%	83.7 (±1.0%)

	Single	Shared	All
3 layer 256 nodes ea.	0.73 MB	2.6 MB	9.2 MB
3 layer 512 nodes ea.	0.80 MB	2.7 MB	9.4 MB
3 layer 1024 nodes ea.	2.92 MB	10.4 MB	36.8 MB



Petko Georgiev, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, "Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations", IMWUT '17

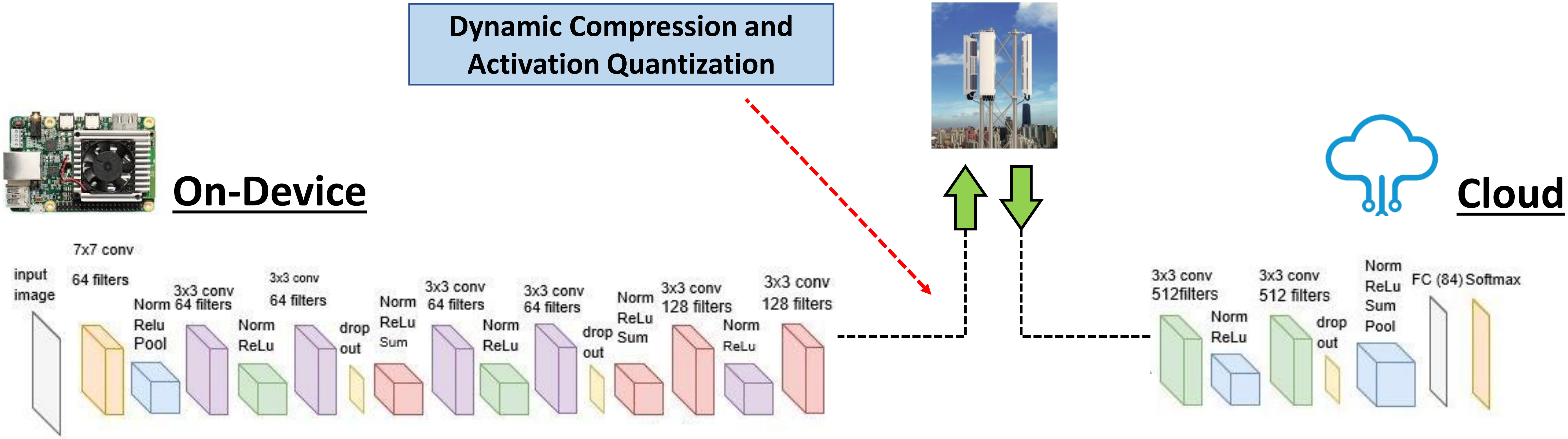
Sharing Resource Example: Exposing Cloud Capacity w/ Device ML by **Dynamic Quantization & Compression**



Decision Factors

- Estimated {Device, Network, Cloud} Latency
- Intensity of Compression and Quantization

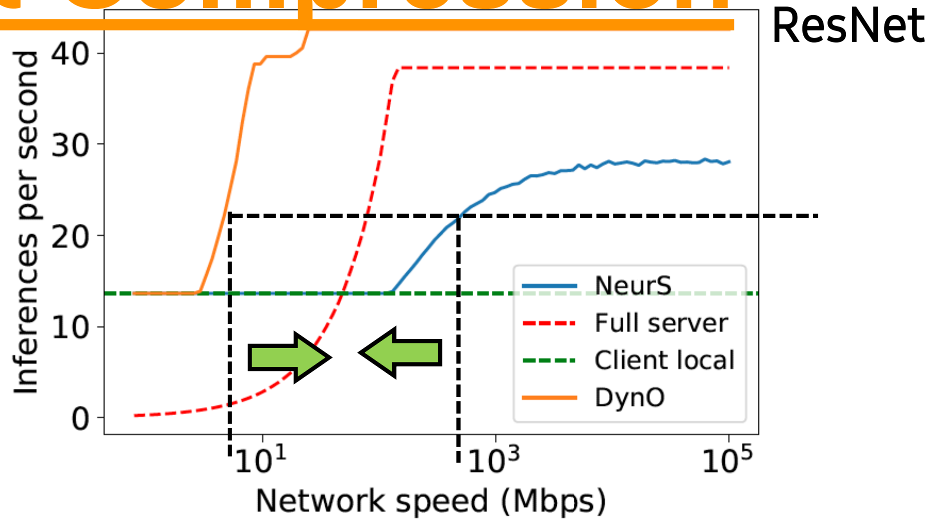
Sharing Resource Example: Exposing Cloud Capacity w/ Device ML by **Dynamic Quantization & Compression**

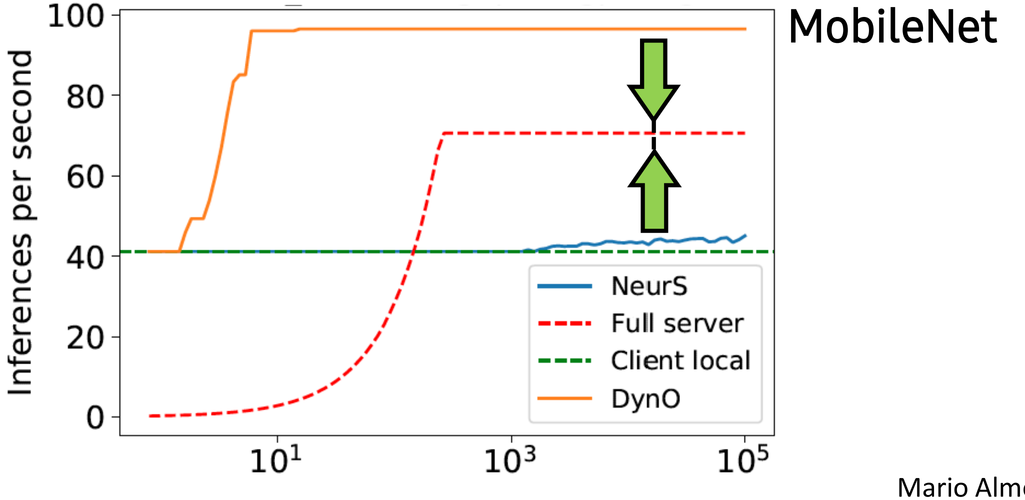


Decision Factors

- Estimated {Device, Network, Cloud} Latency
- Intensity of Compression and Quantization

Sharing Resource Example: Exposing Cloud Capacity w/ Device ML by **Dynamic Quantization & Compression**

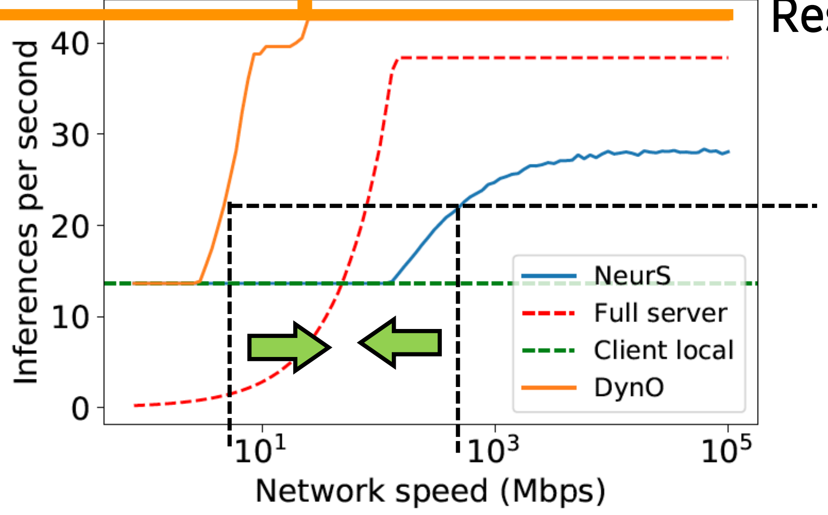




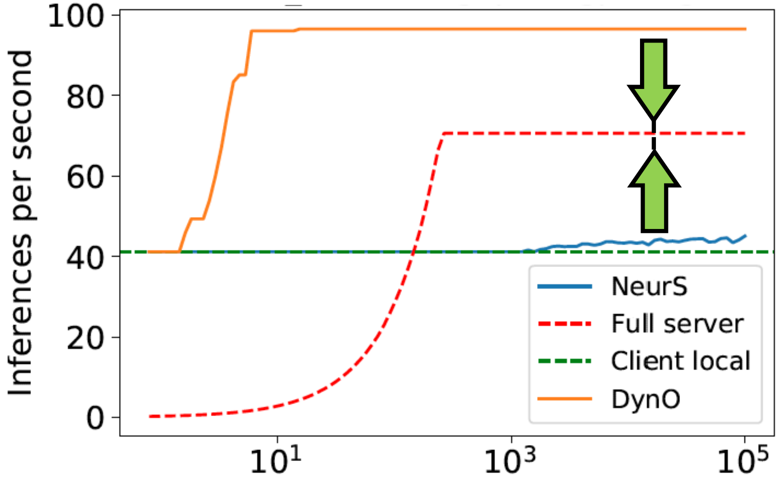
Mario Almeida, Stefanos Laskaridis, Ilias Leontiadis, Stylianos I Venieris, Nicholas D. Lane, "Dyno: Dynamic Onloading of Deep Neural Networks from Cloud to Device", *under submission SysML '20*

Sharing Resource Example: Exposing Cloud Capacity w/ Device ML by **Dynamic Quantization & Compression**

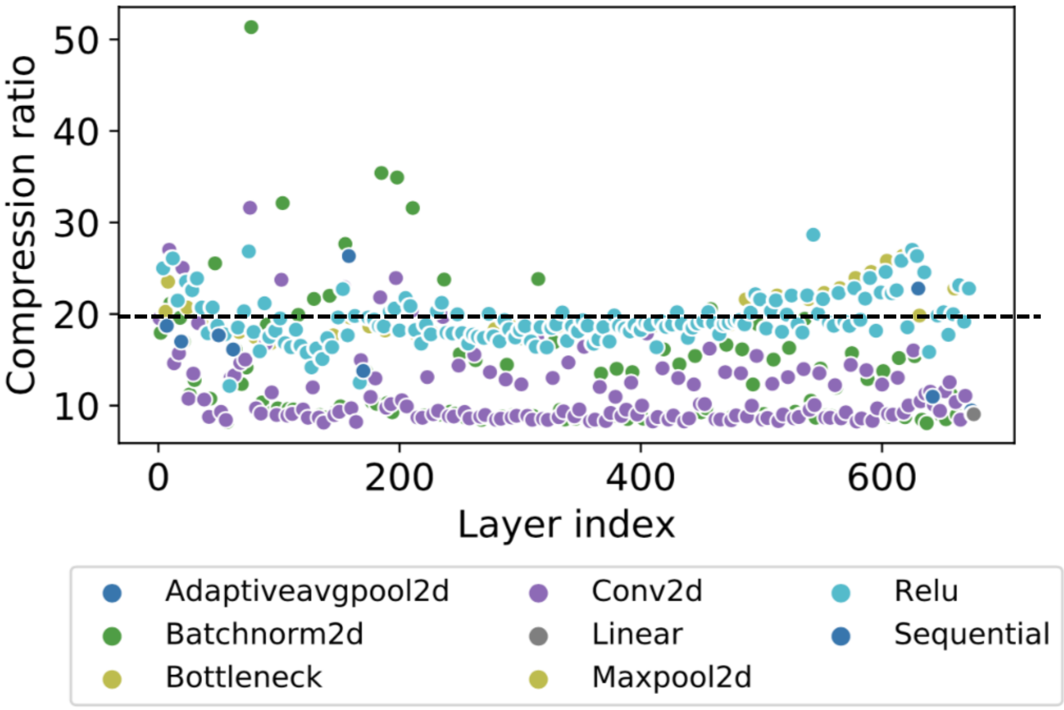
ResNet



MobileNet



20x reduction

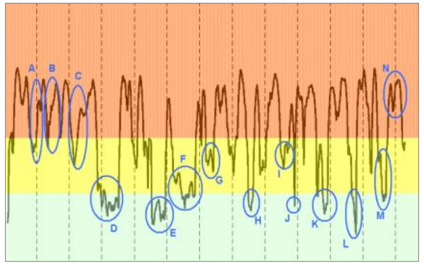


Mario Almeida, Stefanos Laskaridis, Ilias Leontiadis, Stylianos I Venieris, Nicholas D. Lane, "Dyno: Dynamic Onloading of Deep Neural Networks from Cloud to Device", *under submission SysML '20*

Predictions for the ML Efficiency Revolution

#1 Enabling devices to go far beyond classification

#2 Key contributions to the advancement of ML broadly

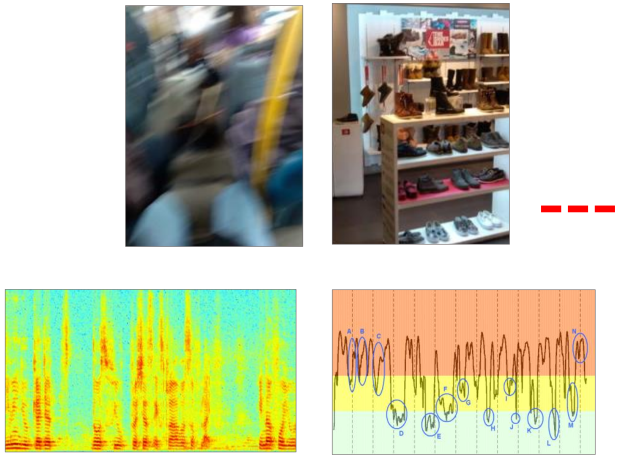


Discriminative Task

{ step count, sleep hours }

#1 ML Efficiency Prediction

On-Device AI goes far beyond classification



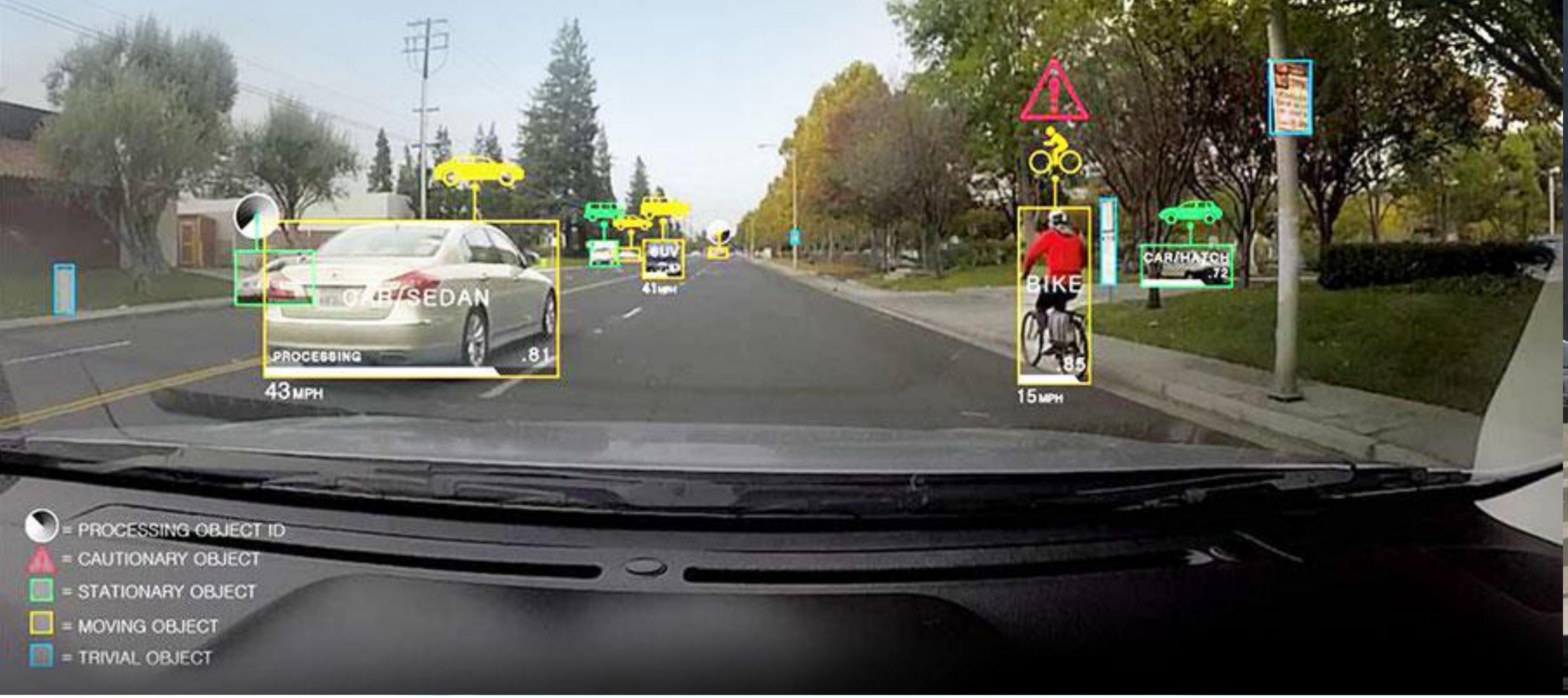
Reasoning

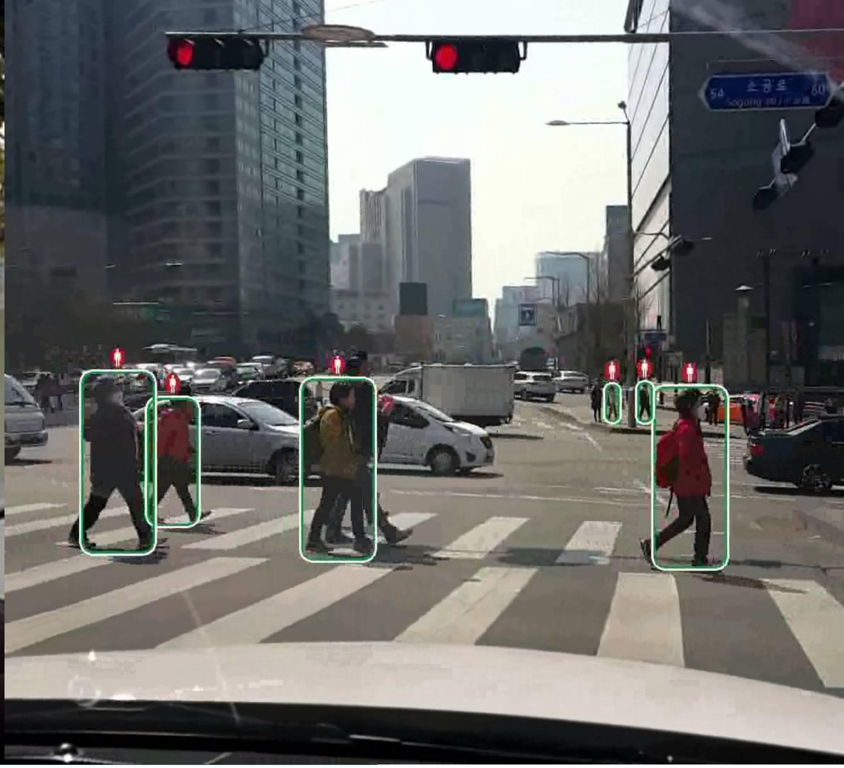
Common Sense

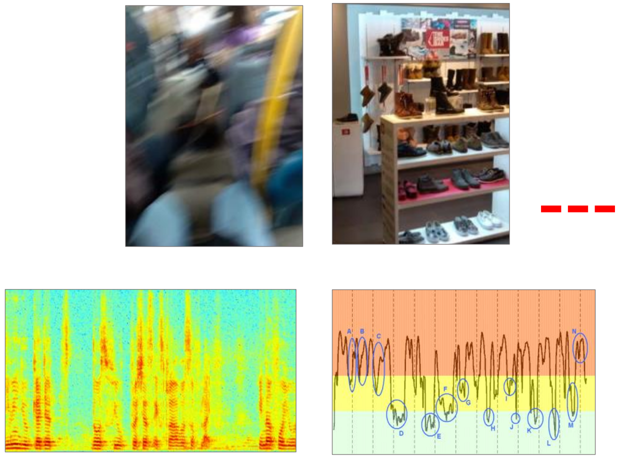
Understanding

Perception *(Discriminative)*

Cognitive Mobile Stack

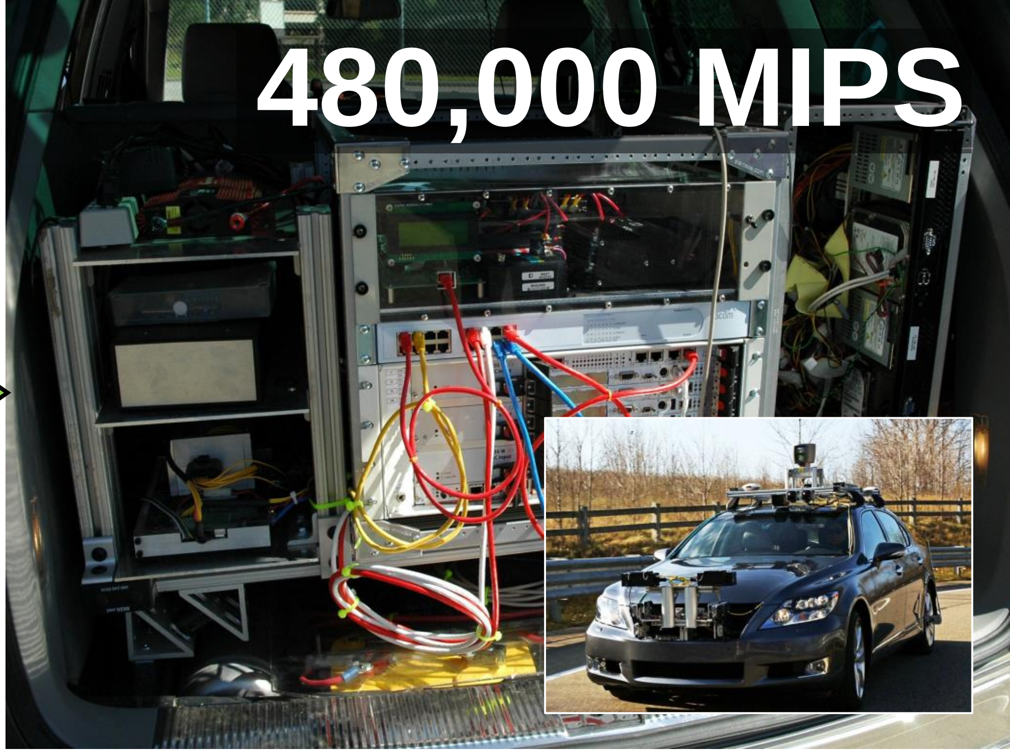


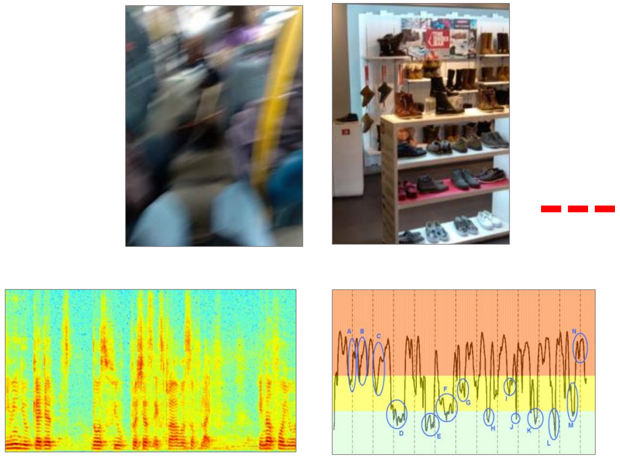




Reasoning
Common Sense
Understanding
Perception *(Discriminative)*

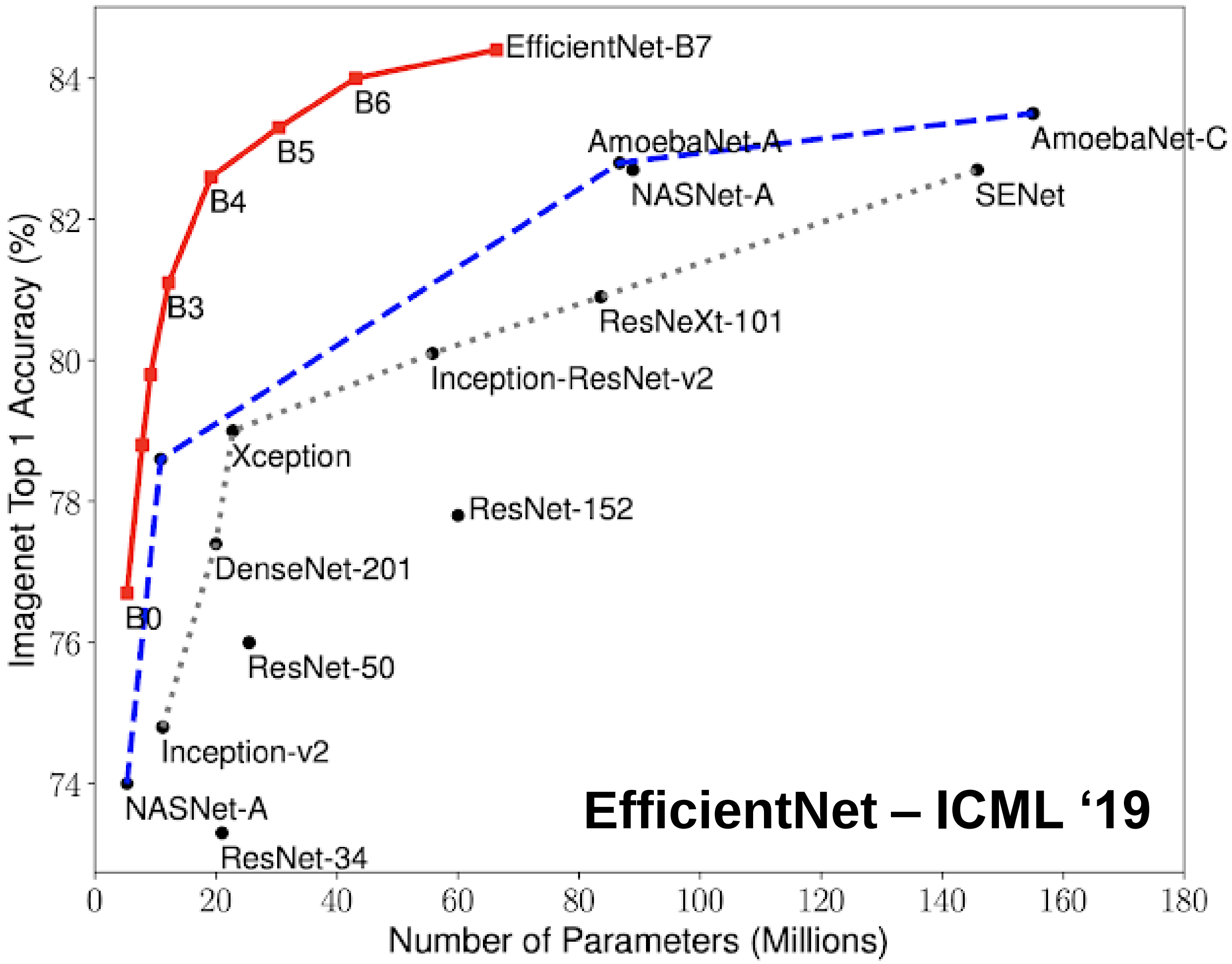
Cognitive Mobile Stack





Reasoning
Common Sense
Understanding
Perception *(Discriminative)*

Cognitive Mobile Stack



Impact of Efficiency

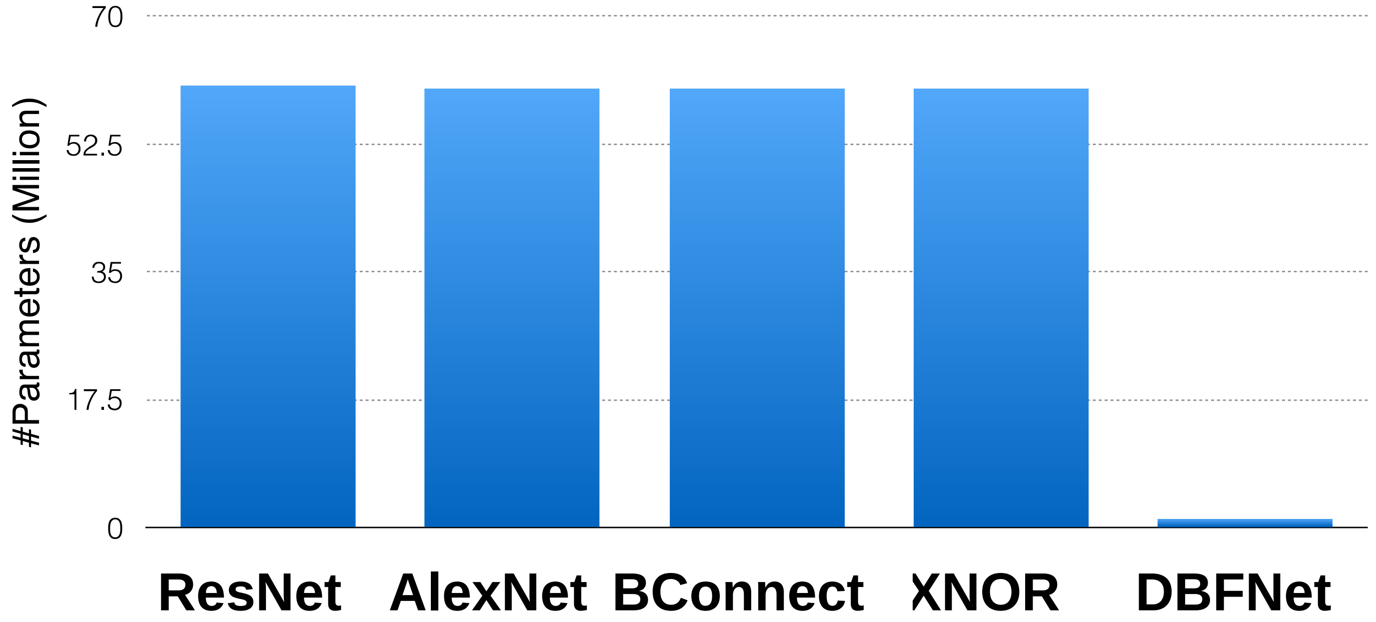
- Faster exploration
- Making feasible powerful *“intractable” approaches*
- More data
- Larger architectures
- New tasks

#2 ML Efficiency Prediction

SOA Accuracy will come from Efficient Models

	ResNet-18	ResNet-34	SqueezeNet
DBF layers	91.15%	92.46%	91.16%
non DBF layers	91.02%	92.36%	91.33%

DBFNet – IJCAI '18



Impact of Efficiency

- Faster exploration
- Making feasible powerful *"intractable" approaches*
- More data
- Larger architectures
- New tasks

#2 ML Efficiency Prediction

SOA Accuracy will come from Efficient Models

Thanks! Questions?

UNIVERSITY OF OXFORD

SAMSUNG

Select Publications

- “An Empirical study of Binary Neural Networks' Optimisation” – ICLR 2019
- “EmBench: Quantifying Performance Variations of Deep Neural Networks across Modern Commodity Devices” – EMDL 2019
- “MobiSR: Efficient On-Device Super-Resolution through Heterogeneous Mobile Processors” – MobiCom 2019
- “Mic2Mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems” – IPSN 2019
- “The deep (learning) transformation of mobile and embedded computing” – IEEE Computer Magazine, 51 (5), 2018
- “BinaryCmd: Keyword Spotting with Deterministic Binary Basis” – SysML 2018
- “Deterministic binary filters for convolutional neural networks” – IJCAI 2018
- “Multimodal Deep Learning for Activity and Context Recognition” – UbiComp 2018
- “Accelerating Mobile Audio Sensing Algorithms through On-Chip GPU Offloading” – MobiSys 2017
- “Squeezing Deep Learning into Mobile and Embedded Devices” – IEEE Pervasive Magazine, 16 (3), 2017
- “Cross-modal recurrent models for weight objective prediction from multimodal time-series data” – *Pervasive Health 2018*
- “Low-resource Multi-task Audio Sensing for Mobile and Embedded Devices via Shared Deep Neural Network Representations” – UbiComp 2017
- “DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models using Wearable Commodity Hardware” – MobiSys 2017
- “Sparsifying Deep Learning Layers for Constrained Resource Inference on Wearables” – SenSys 2016
- “X-CNN: Cross-modal convolutional neural networks for sparse datasets” – SSCI 2016
- “DXTK: Enabling resource-efficient deep learning on mobile and embedded devices with the deepx toolkit” – MobiCASE 2016
- “LEO: Scheduling sensor inference algorithms across heterogeneous mobile processors and network resources” – MobiCom 2016
- “From Smart to Deep: Robust Activity Recognition on Smartwatches using Deep Learning” – WristSense 2016
- “Deepx: A software accelerator for low-power deep learning inference on mobile devices”— IPSN 2016
- “An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices” – IoTApp 2015
- “Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning” – UbiComp 2015
- “Can Deep Learning Revolutionize Mobile Sensing?” – HotMobile 2015

Nicholas D. Lane

@niclane7
http://mlsys.cs.ox.ac.uk