UNIVERSITY OF CAMBRIDGE
Cambridge Machine Learning Systems Lab
http://mlsys.cst.cam.ac.uk

SAMSUNG
Samsung AI
Cambridge, UK

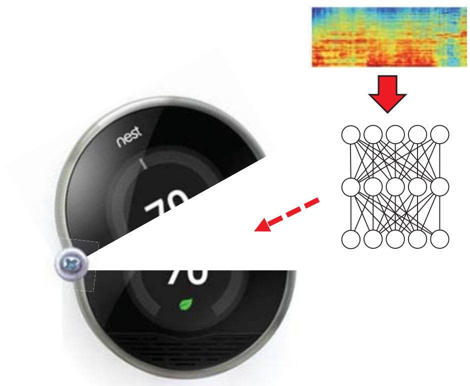
Machine Learning Systems: On-Device AI and Beyond

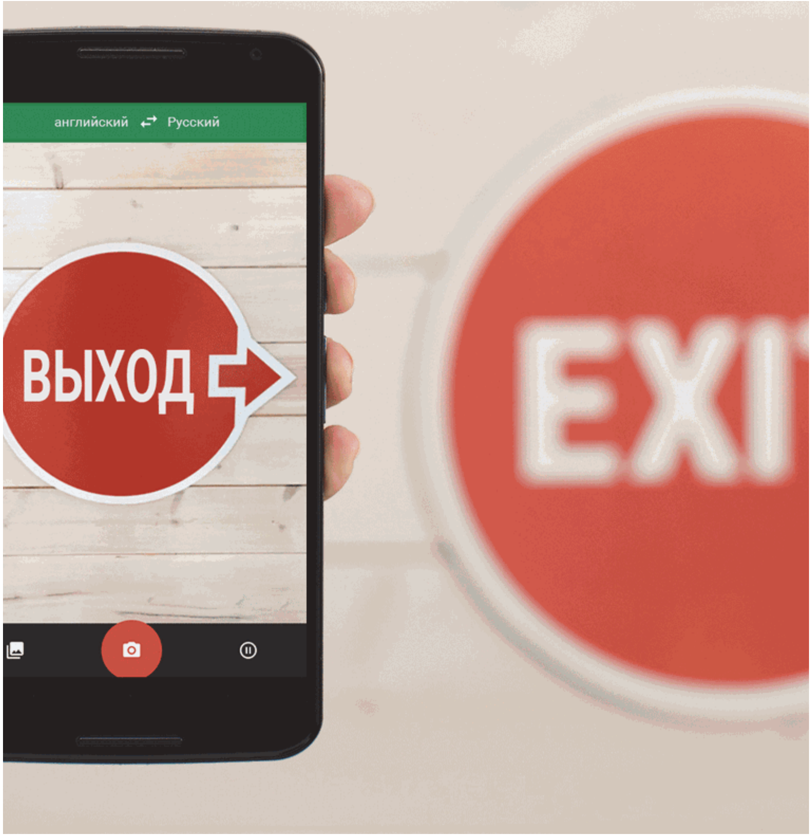
Nicholas D. Lane

TECoSA Seminar Series
Thursday December 3rd 2020

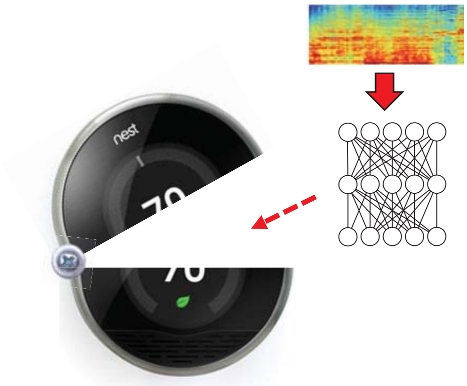
@niclane7 | http://niclane.org
ndl32@cam.ac.uk

Efficient ML Revolution

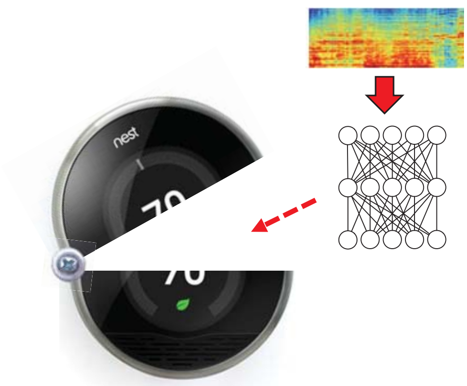




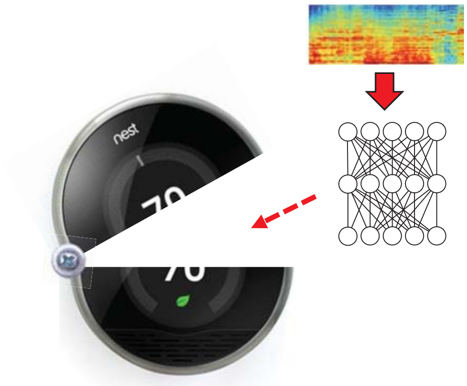
Efficient ML Revolution



Efficient ML Revolution

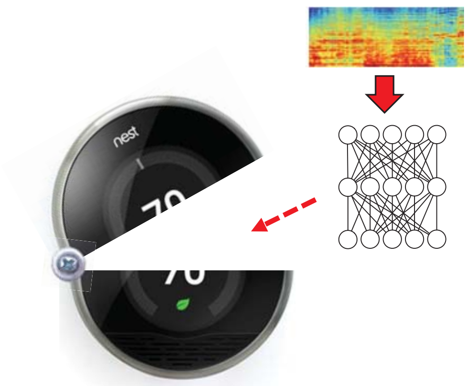


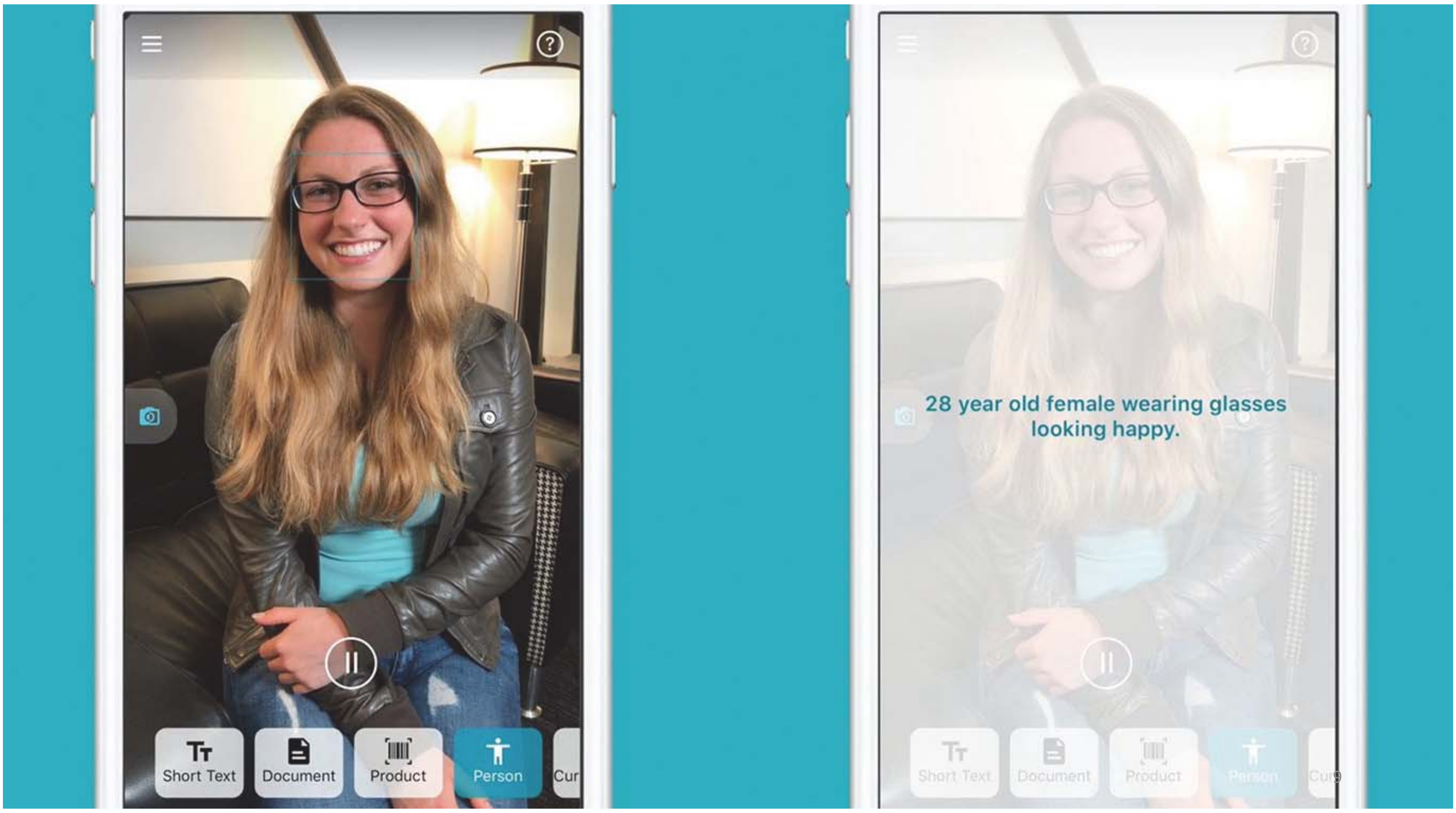
Efficient ML Revolution



- User Privacy
- Strongest algorithms for the hardest problems
- Robust devices without a network dependency
- Low energy and latency

Efficient ML Revolution





From Late '14: Efficient ML Revolution

- **2014** — 1st Proof-of-Concept DL of Audio/IMU on Mobile [HotMobile '15]
- **2015** — DeepEar (1st DSP-based DL General Audio Sensing) [UbiComp '15]
- **2016** — DL Smartwatch Activity Recognition [WristSense '16]
- 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]

Community Innovations

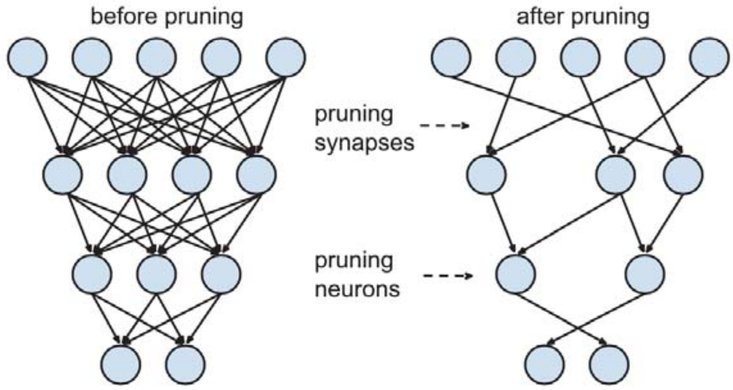
Algorithmic & Architecture Advances

- Node Pruning, Leverage Sparsity
- SqueezeNet (50x AlexNet reduction)
- Low Precision (8/4 bit), Binarization
- MobileNet, MCDNN, Custom Nets

Hardware Innovations

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

Model Compression Example: Node Pruning



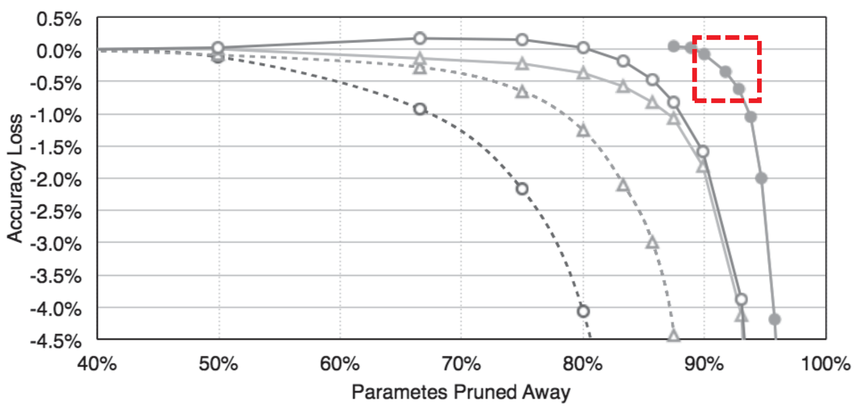
Related Methodologies

- Leveraging Sparsity
- Low Precision Results (8-bit etc.)
- Binarization of Networks
- MobileNet, MCDNN, Custom Nets

.. and *even hardware approaches*

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

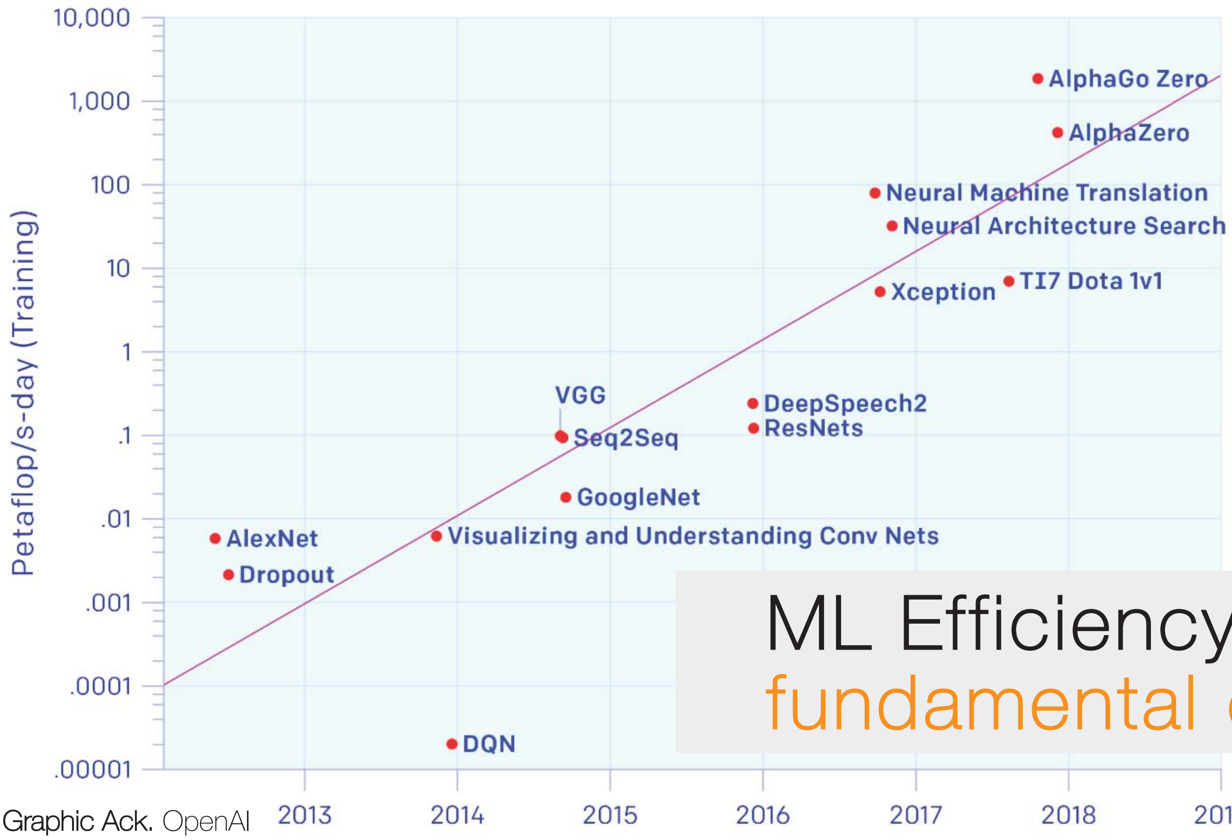
Model Compression Example: Node Pruning



~50x gains

tiny accuracy loss

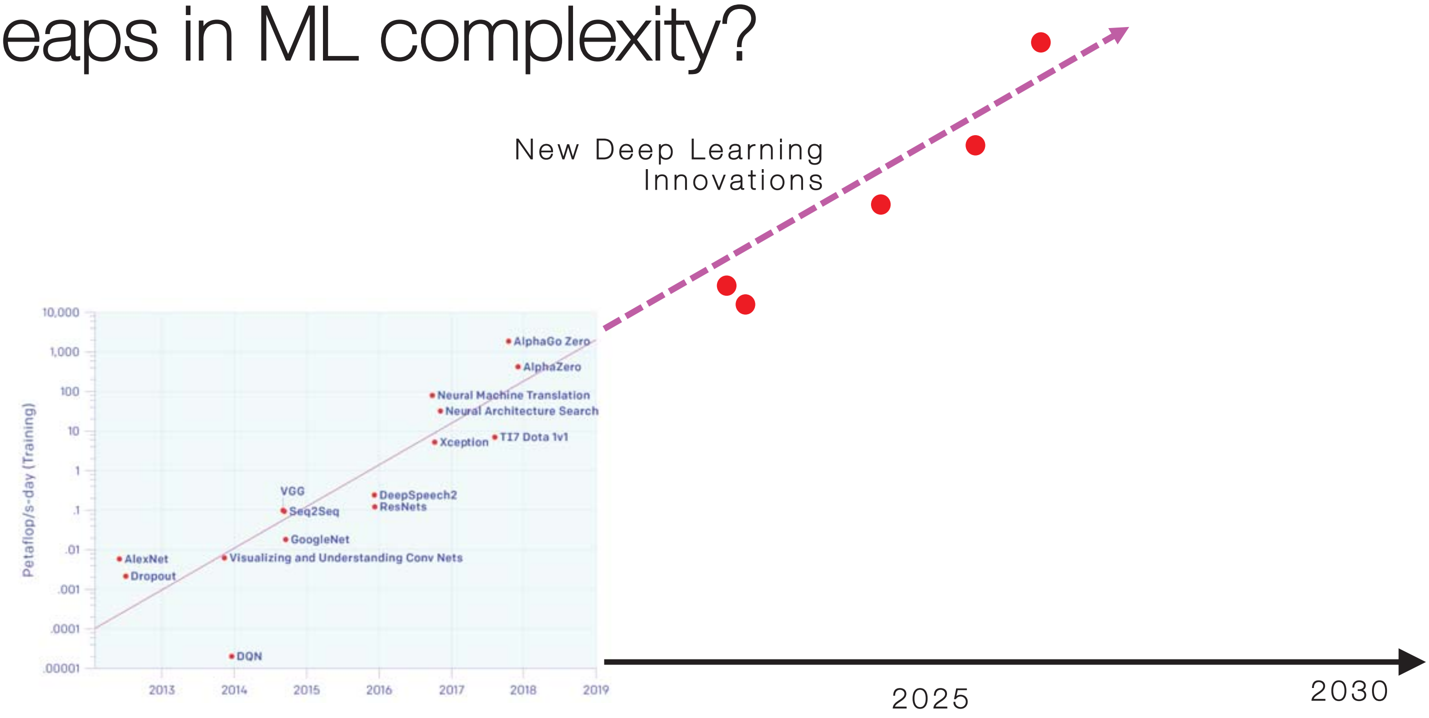
Song Han, Jeff Pool, John Tran, William J. Dally, "Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015



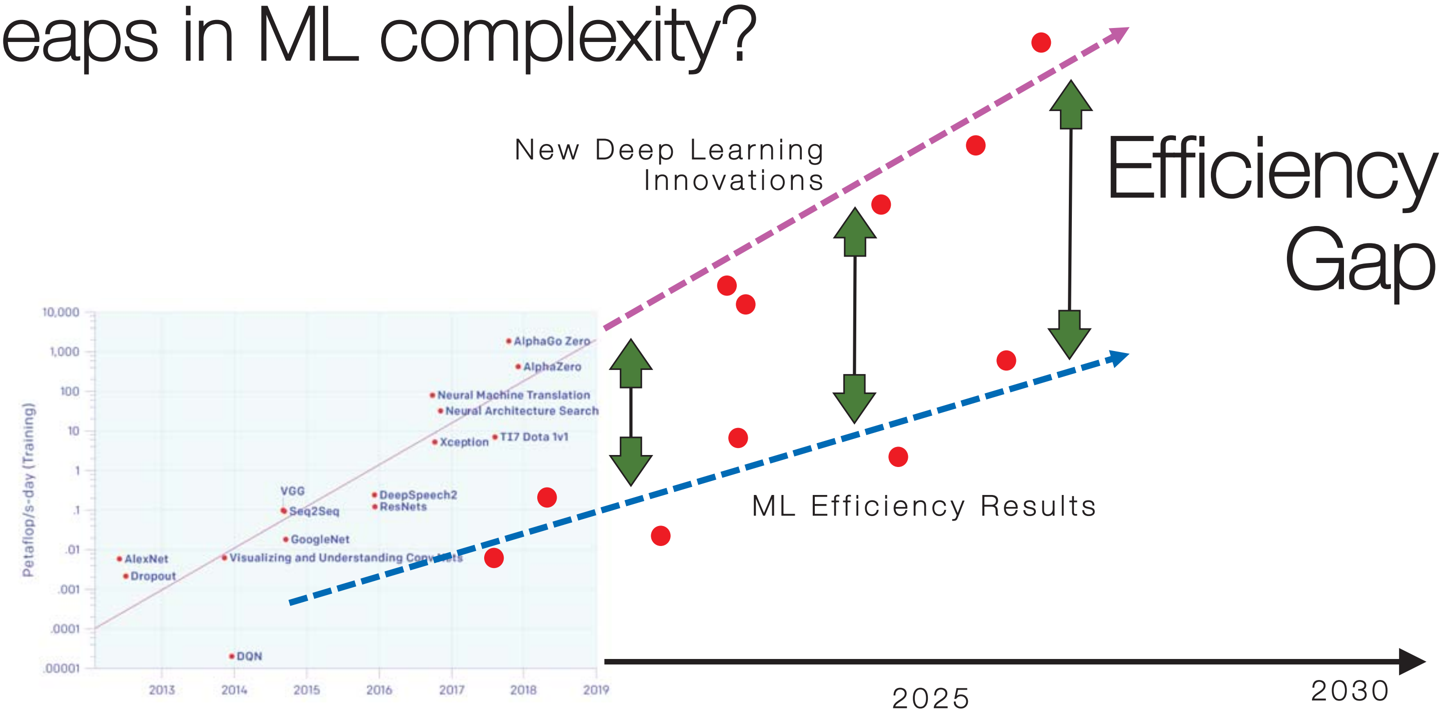
Graphic Ack. OpenAI

ML Efficiency is a
fundamental crisis

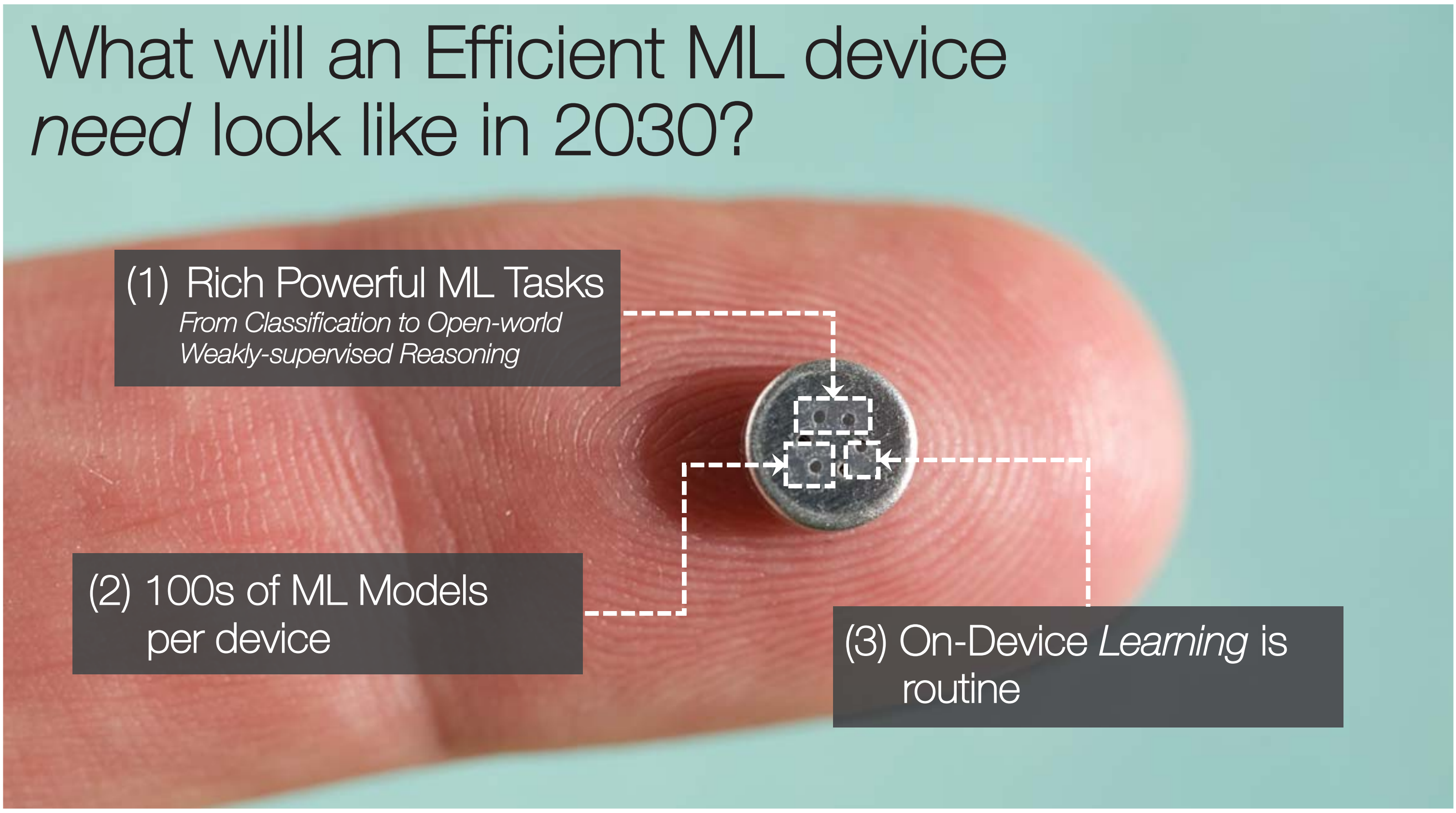
Can we keep up with the upcoming leaps in ML complexity?



Can we keep up with the upcoming leaps in ML complexity?




What will an Efficient ML device
need look like in 2030?

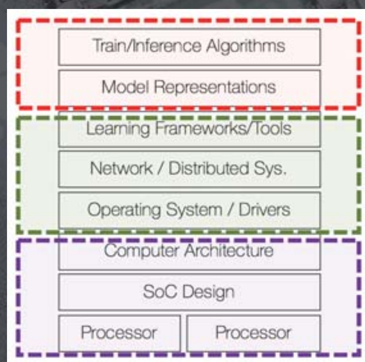

What will an Efficient ML device
need look like in 2030?
(1) Rich Powerful ML Tasks
From Classification to Open-world Weakly-supervised Reasoning
(2) 100s of ML Models per device
(3) On-Device *Learning* is routine

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

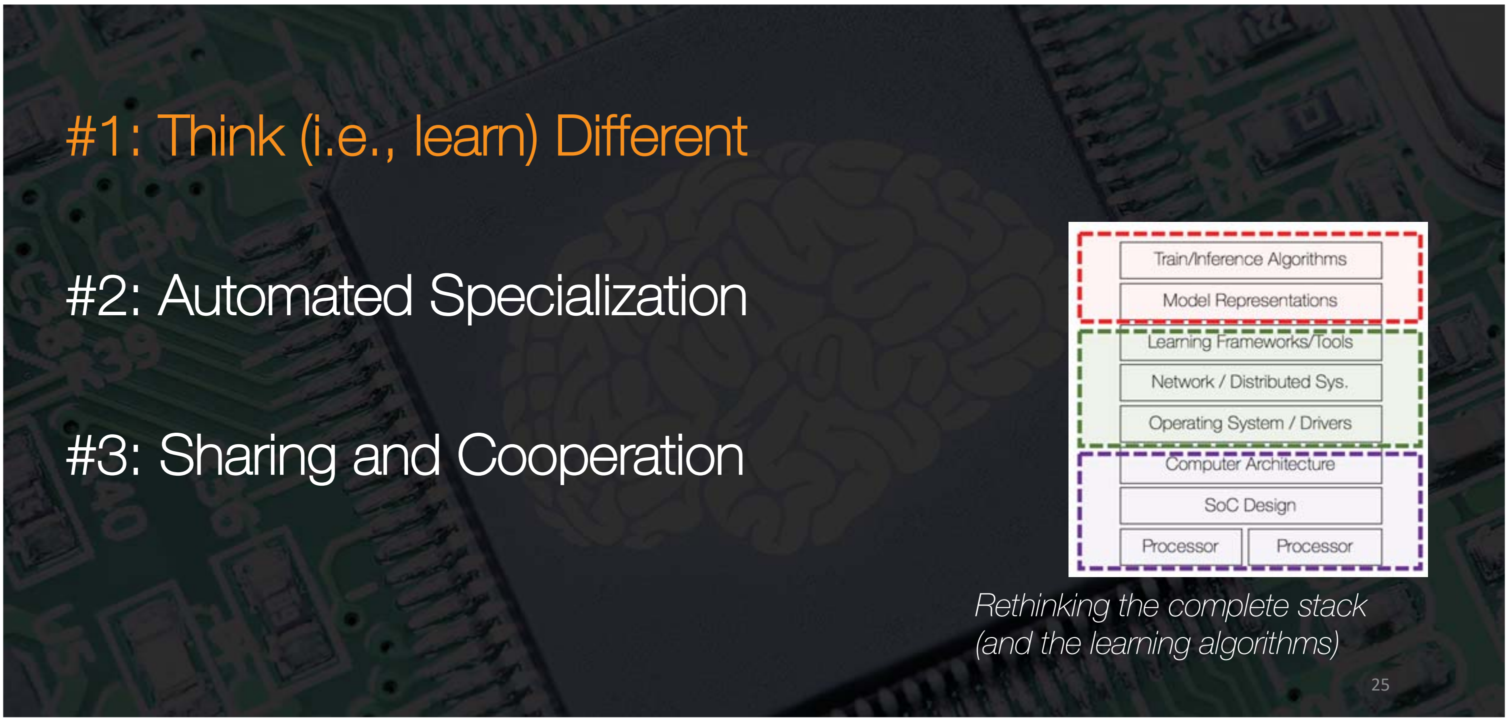
#2: Automated Specialization

#3: Sharing and Cooperation

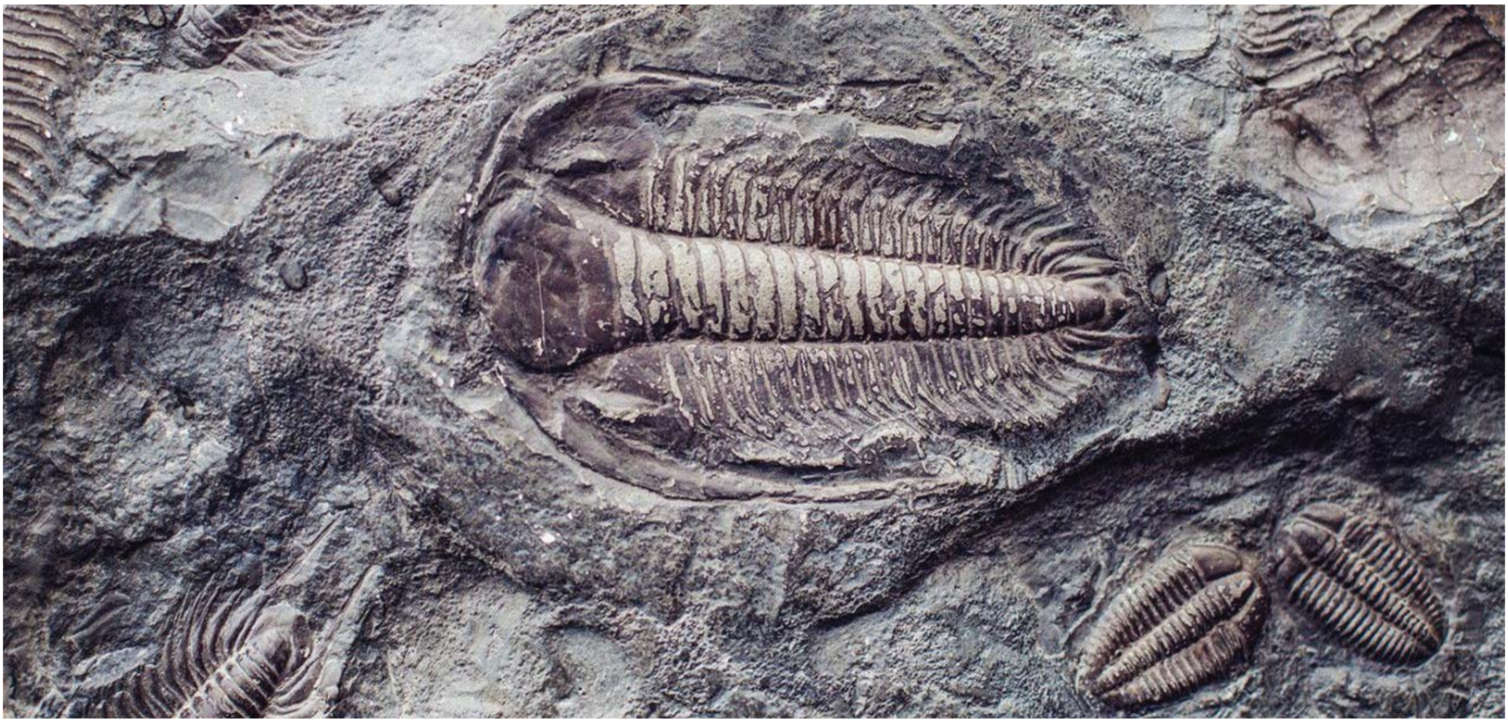


Rethinking the complete stack (and the learning algorithms)

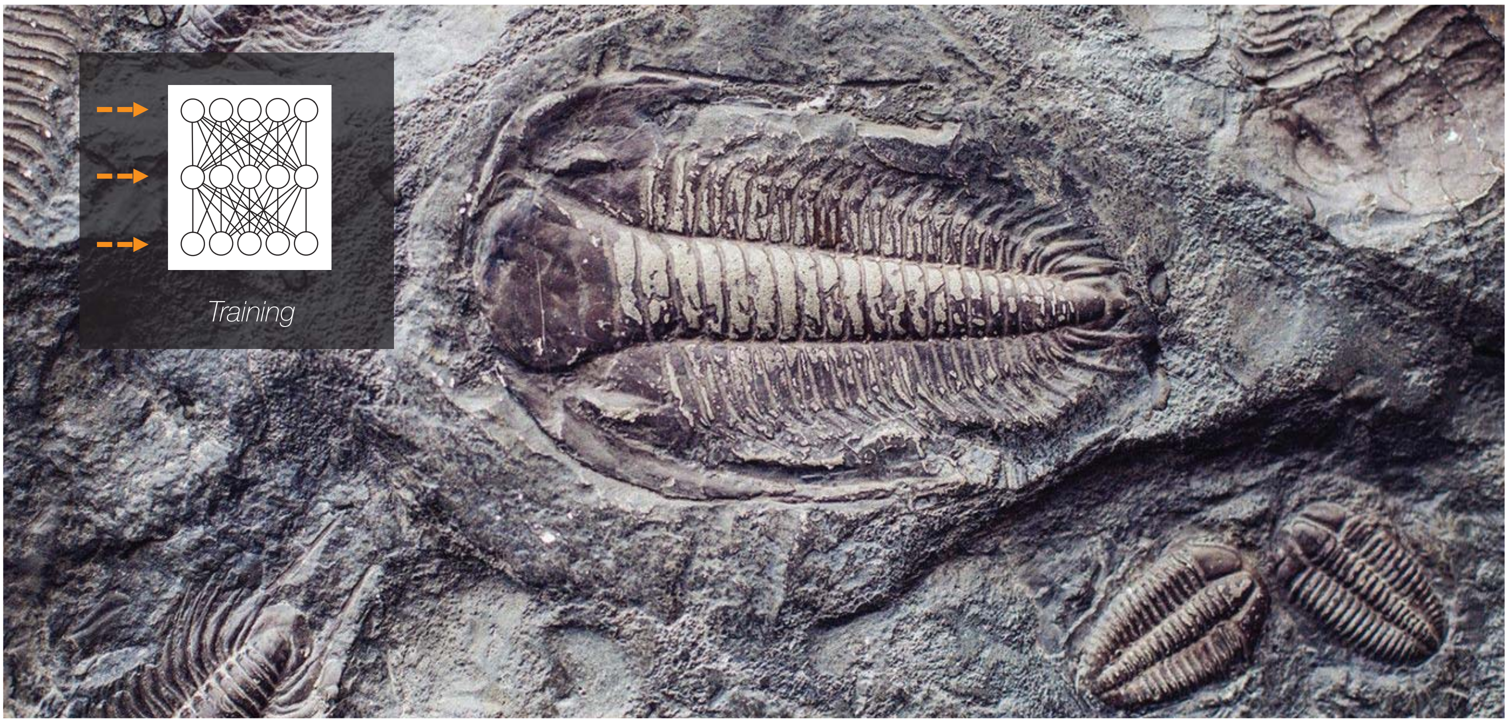
Fundamental Efficient ML Challenges



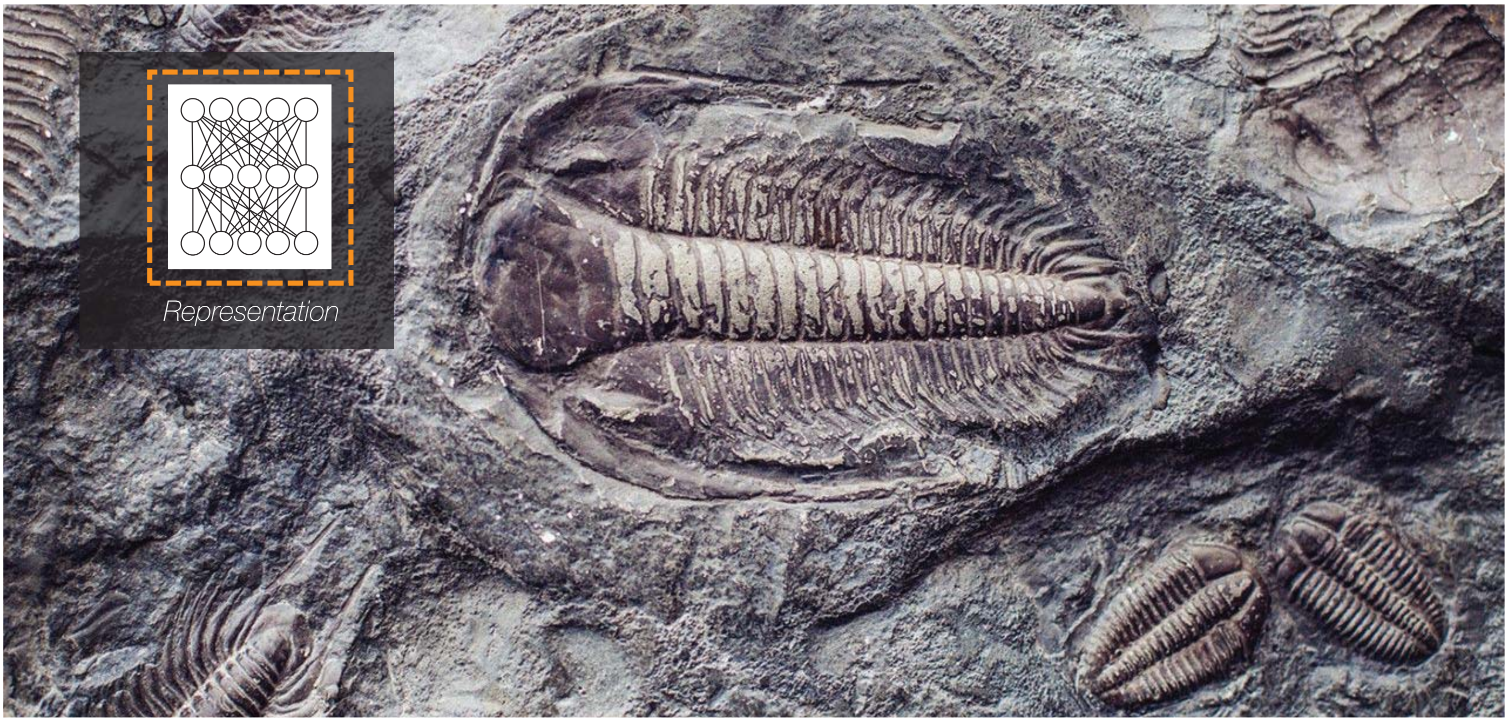
#1 Think (i.e., learn) Different



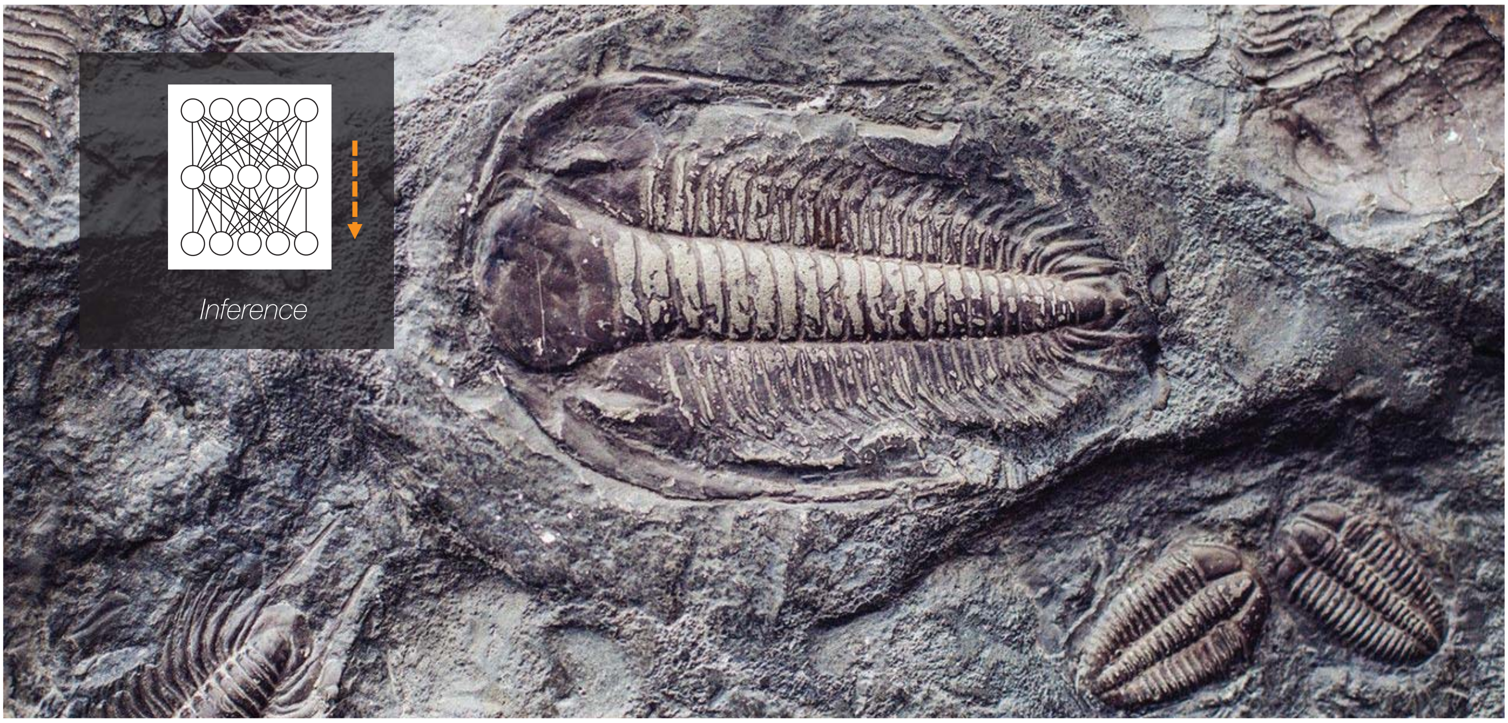
#1 Think (i.e., learn) Different



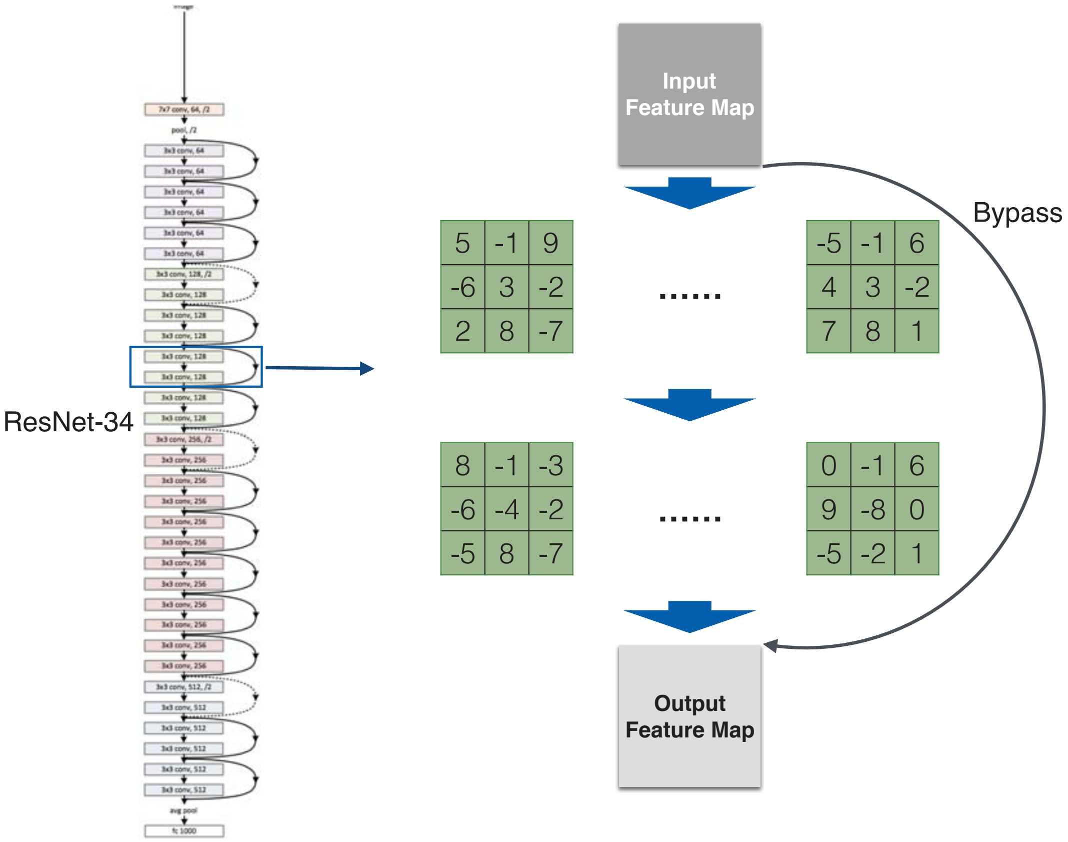
#1 Think (i.e., learn) Different



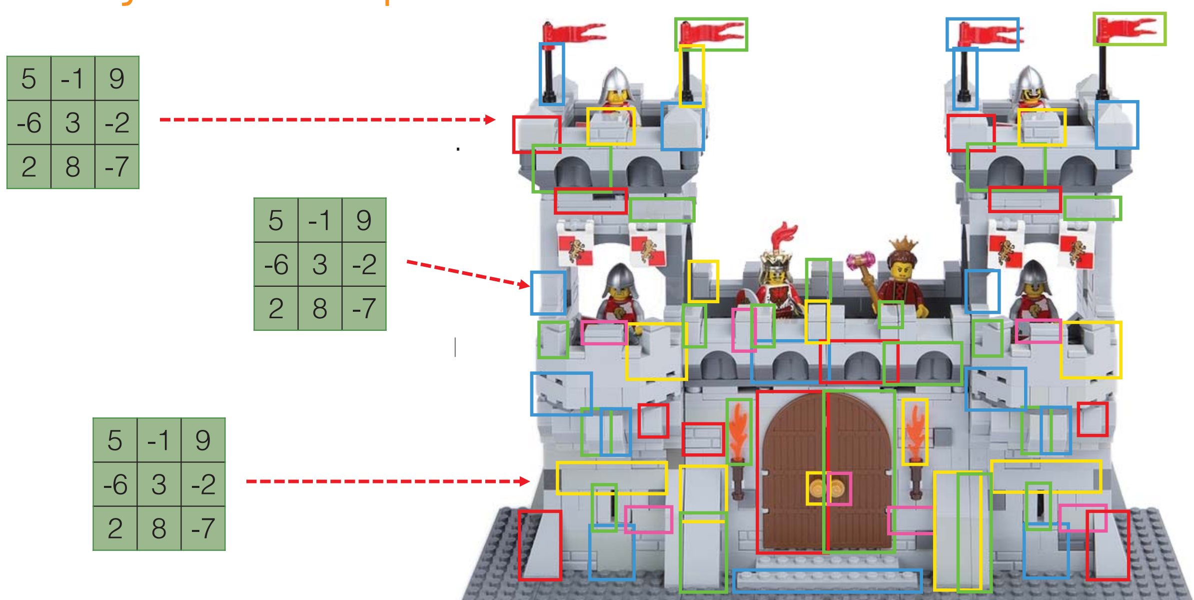
#1 Think (i.e., learn) Different



#1 Think (i.e., learn) Different



Inflating deep models from functions at inference: a new form of trading memory for compute



[IJCAI ‘18, SysML ’18, *submitted* FPL ‘20]

Inflating deep models from functions at inference: a new form of trading memory for compute

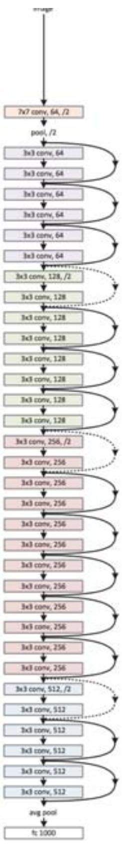
$$f(x,y,z) \simeq \begin{bmatrix} 5 & -1 & 9 \\ -6 & 3 & -2 \\ 2 & 8 & -7 \end{bmatrix}$$

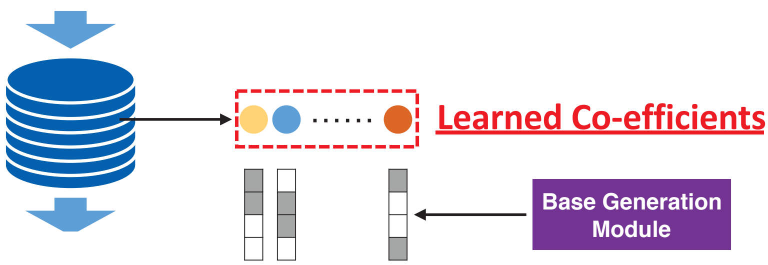
[IJCAI '18, SysML '18, *submitted* FPL '20]

Inference is radically different..

3	8	21	-2
-27	-17	28	24
18	13	-22	-5
-9	6	12	25

ResNet-34



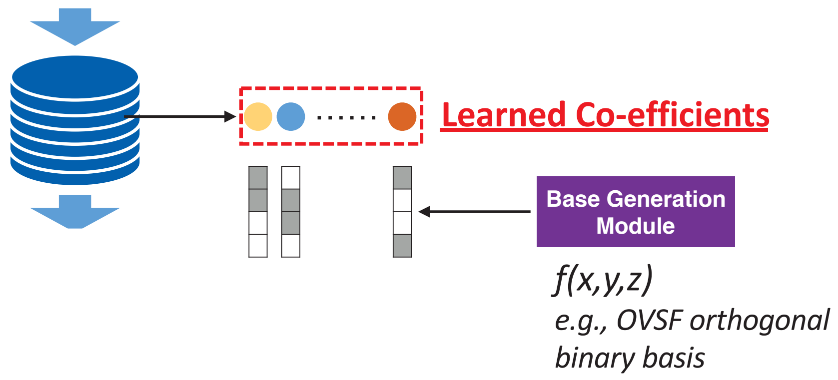


f(x,y,z)
e.g., OVSF orthogonal binary basis

Inference is radically different..

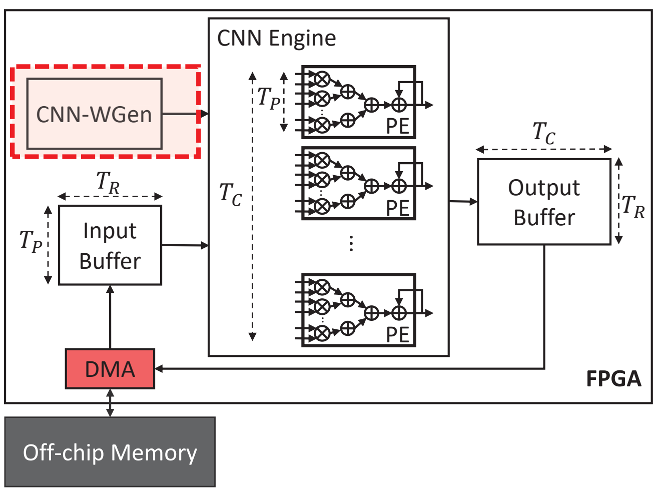
	Top1 Accuracy (%)
ResNet-18	91.15
ResNet-18 (OVSF)	91.02
ResNet-34	92.46
ResNet-34 (OVSF)	92.32
SqueezeNet	91.16
SqueezeNet (OVSF)	91.33

Dataset: CIFAR-10



[IJCAI '18, SysML '18, *submitted* FPL '20]

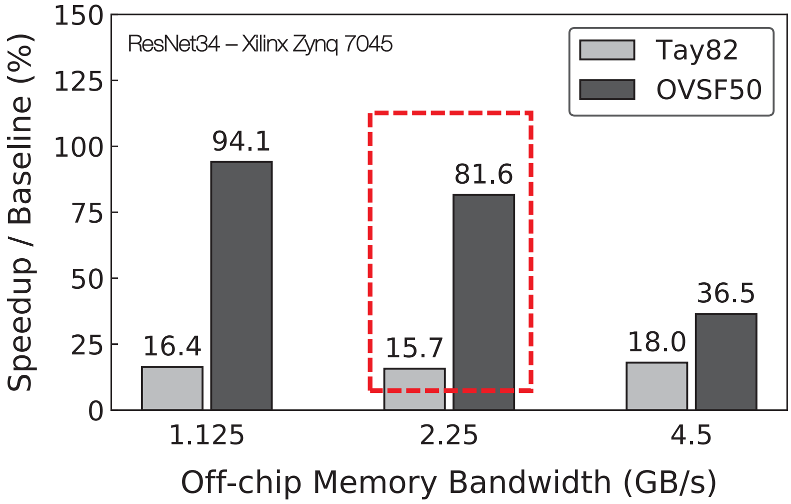
Maximizing efficiency potential through hardware



*unzip*FPGA

[IJCAI '18, SysML '18, *submitted* FPL '20]

Maximizing efficiency potential through hardware



Dataset: ImageNet

Compression (ResNet34)	Params (millions)	Accuracy (%)
None	21.8	73.3
Tay82	17.4	72.7
OVSF50	17.2	72.1

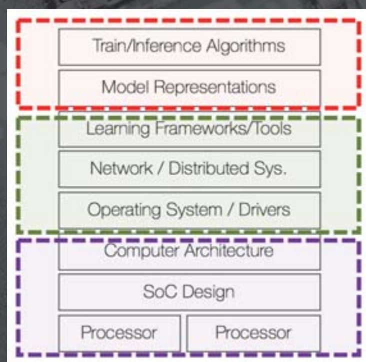
[IJCAI '18, SysML '18, *submitted* FPL '20]

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

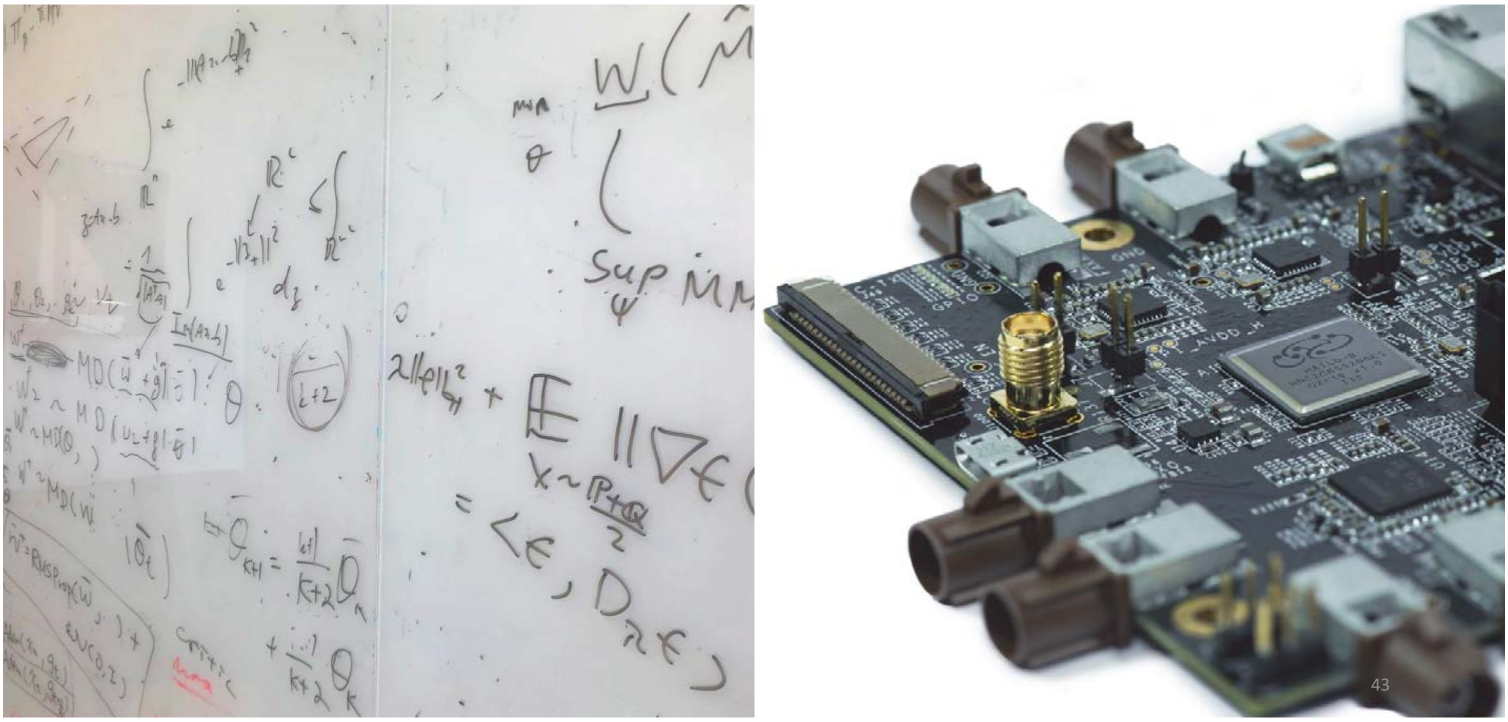
#2: Automated Specialization

#3: Sharing Memory and Compute

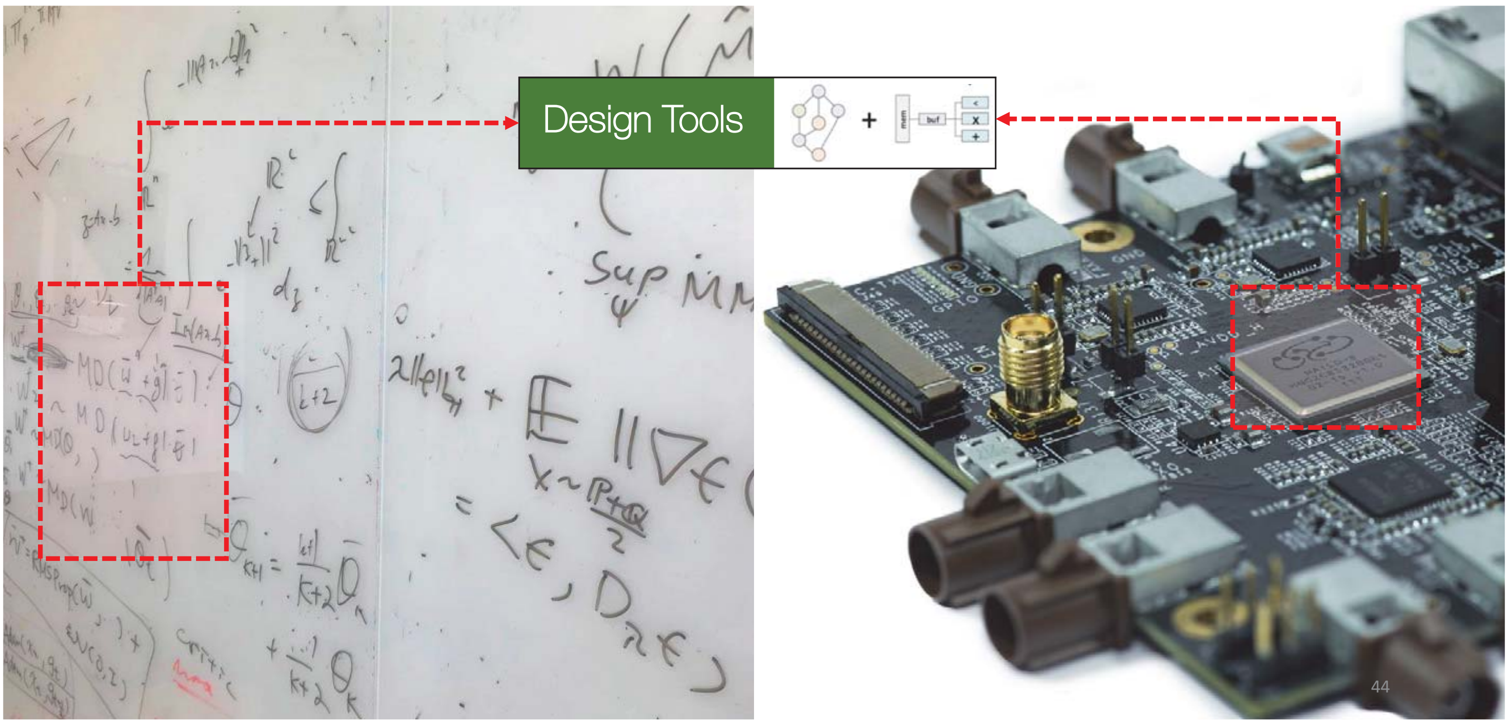


Rethinking the complete stack (and the learning algorithms)

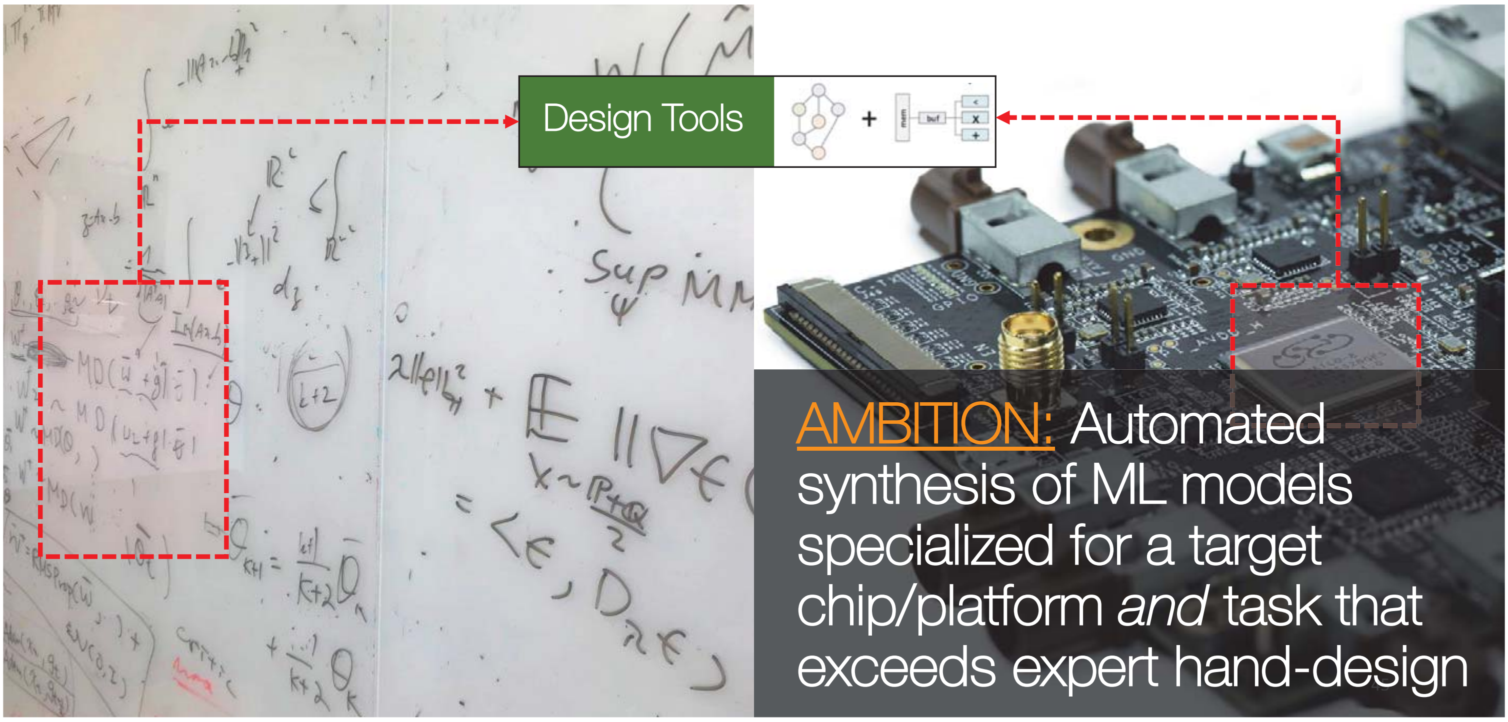
#2 Automated Specialization



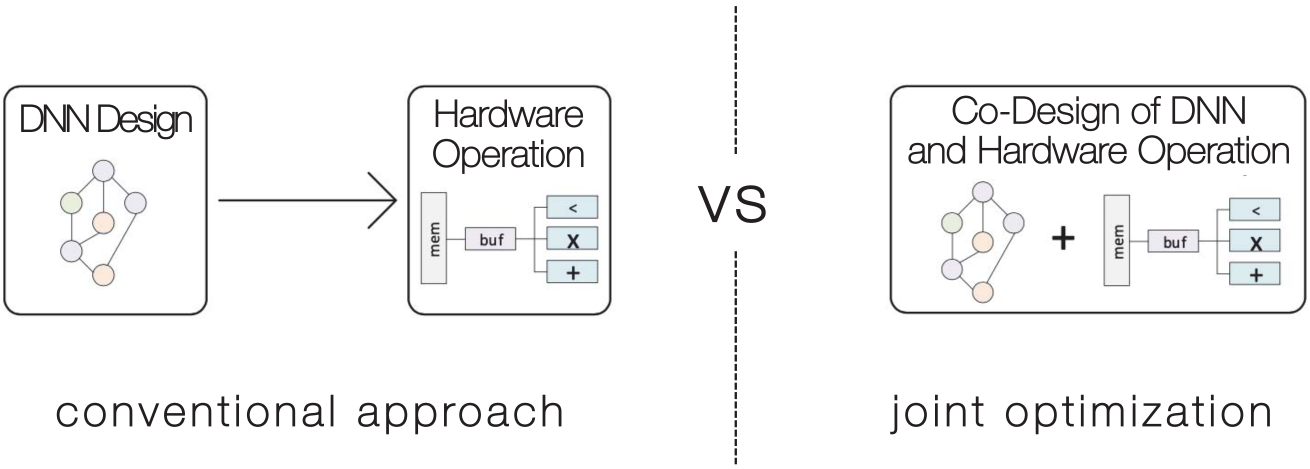
#2 Automated Specialization



#2 Automated Specialization



Joint Optimization of Hardware Operation and Neural Architecture

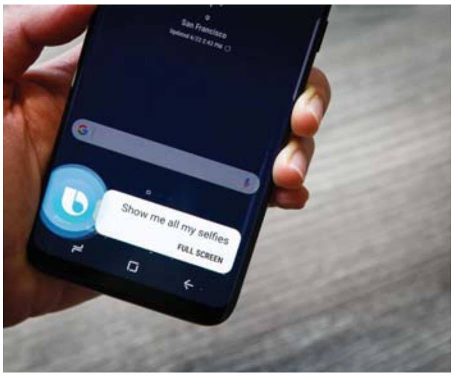


[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Joint Optimization of Hardware Operation and Neural Architecture

Dataset: Librispeech

	WER (test-clean)	Speedup
MoCHA	8.32	1x
AutoCAML	<u>8.19</u>	<u>5.1x</u>



[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Joint Optimization of Hardware Operation and Neural Architecture

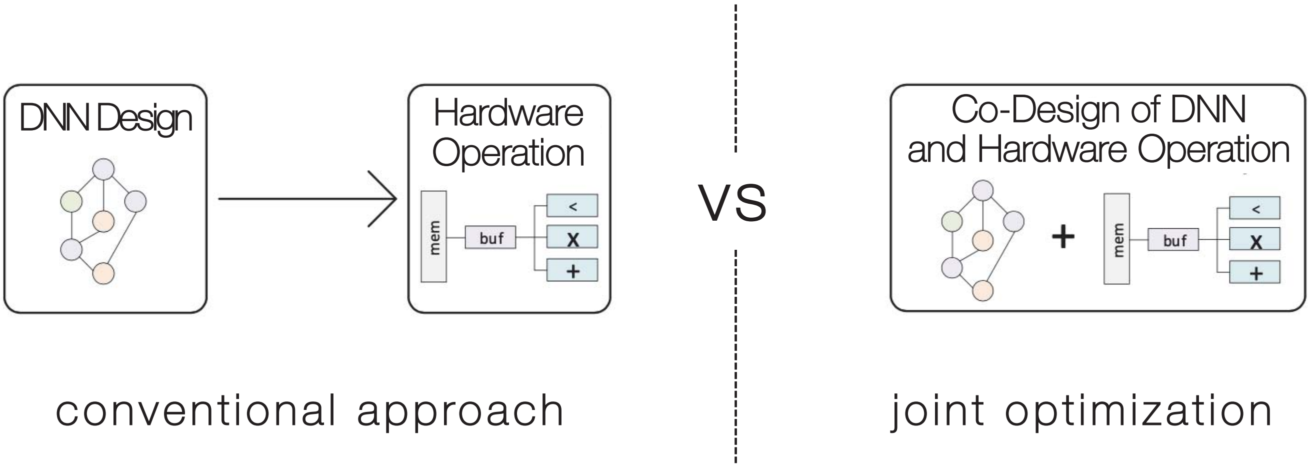
Dataset: Set5

	LPIPS	Mult-Adds (G)	Params (K)
ESRGAN	0.074	1034.1	16,697
FEQE	0.091	5.6	96
AutoCAML	<u>0.076</u>	<u>3.6</u>	<u>61</u>

AutoCAML (61k) ESRGAN (16,697k)

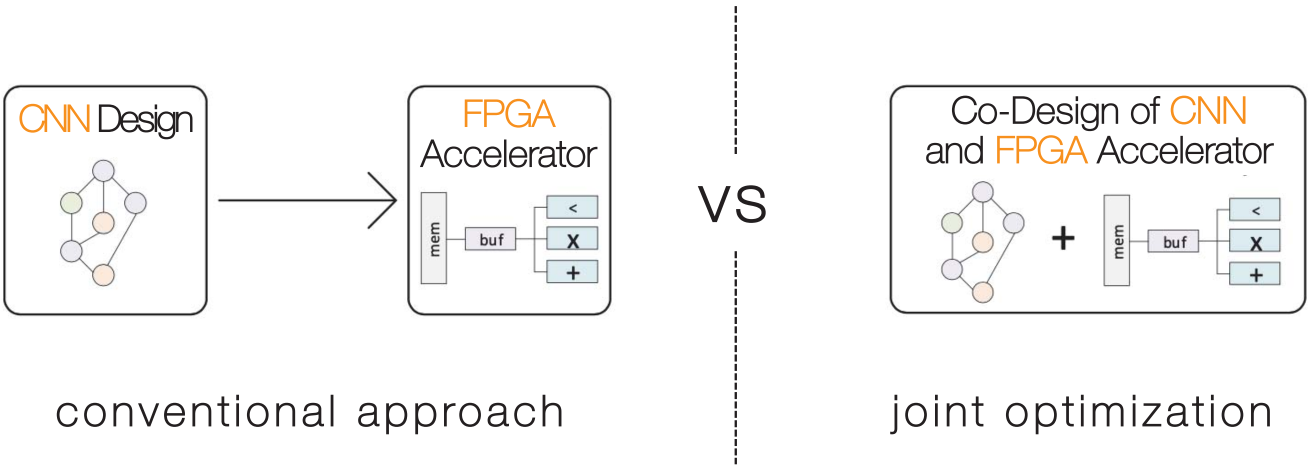
[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Joint Optimization of Hardware Operation and Neural Architecture



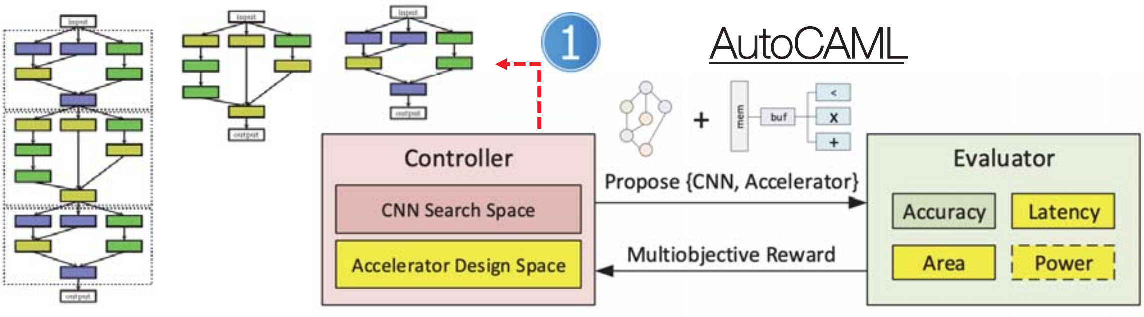
[INTERSPEECH ‘19, INTERSPEECH ’20, EECV ’20, DAC ’20, NeurIPS ‘20]

Joint Optimization of Hardware Operation and Neural Architecture



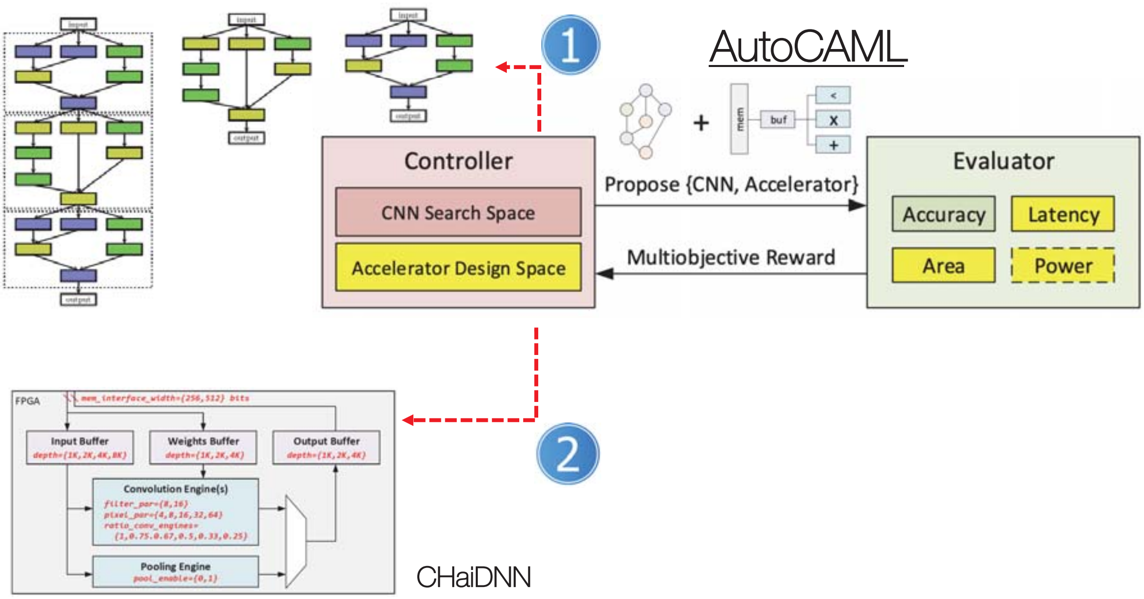
[INTERSPEECH ‘19, INTERSPEECH ’20, EECV ’20, DAC ’20, NeurIPS ‘20]

Joint Optimization of Hardware Operation and Neural Architecture



[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

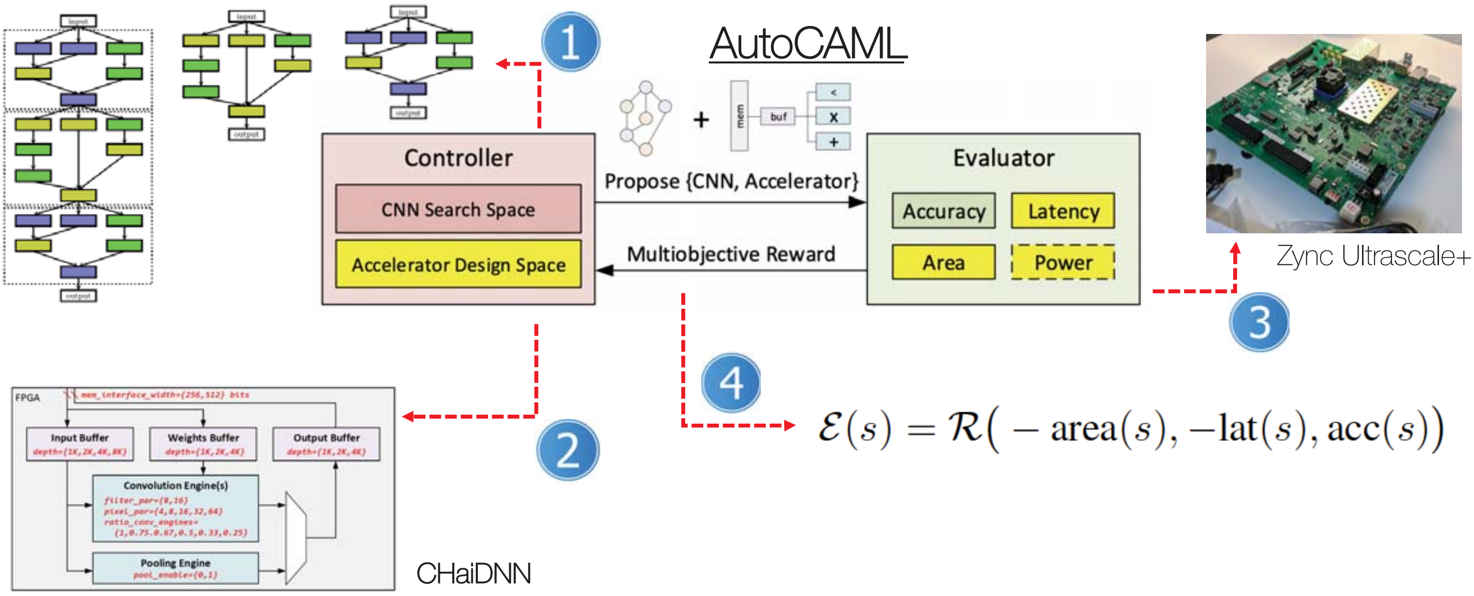
Joint Optimization of Hardware Operation and Neural Architecture



CHaiDNN

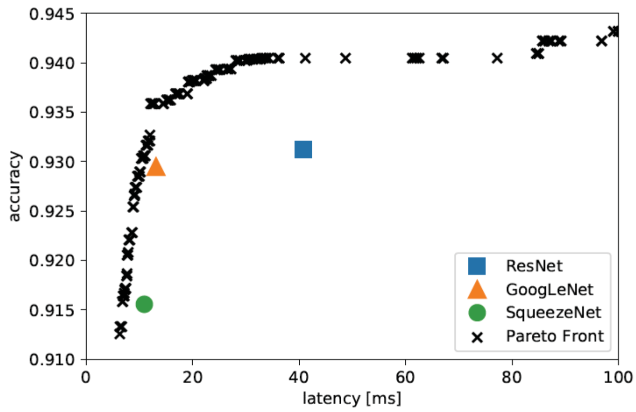
[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

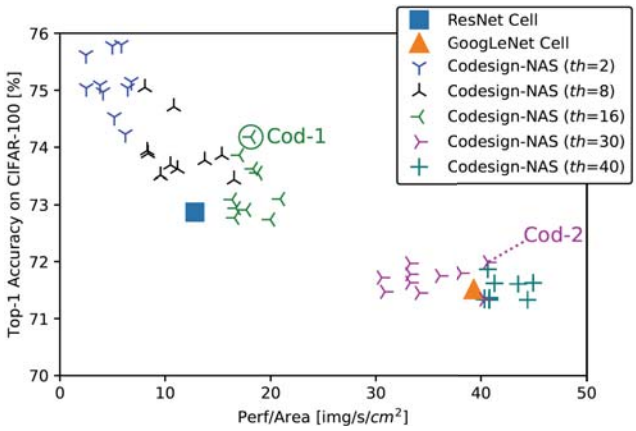
Joint Optimization of Hardware Operation and Neural Architecture



[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Joint Optimization of Hardware Operation and Neural Architecture



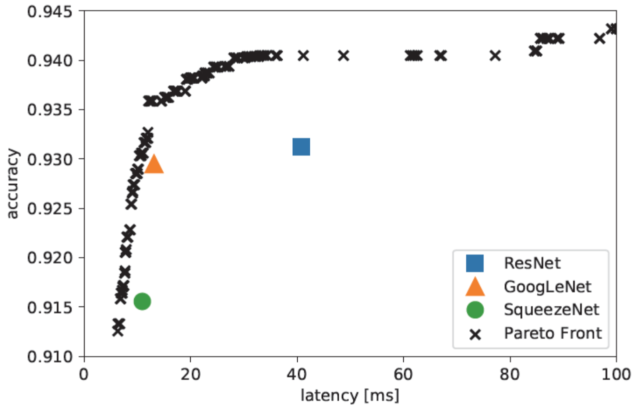


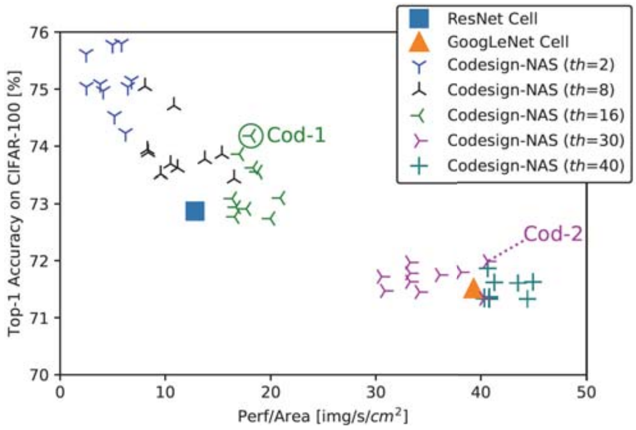
[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Joint Optimization of Hardware Operation and Neural Architecture

	prior SOA	AutoCAML
Accuracy	92.8%	93.6%
Latency	51ms	42ms
HW Area	170mm^2	130mm^2

Zync Ultrascale+





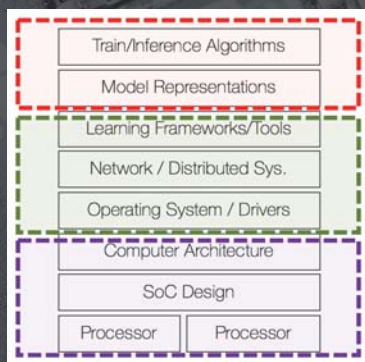
[INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20]

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation

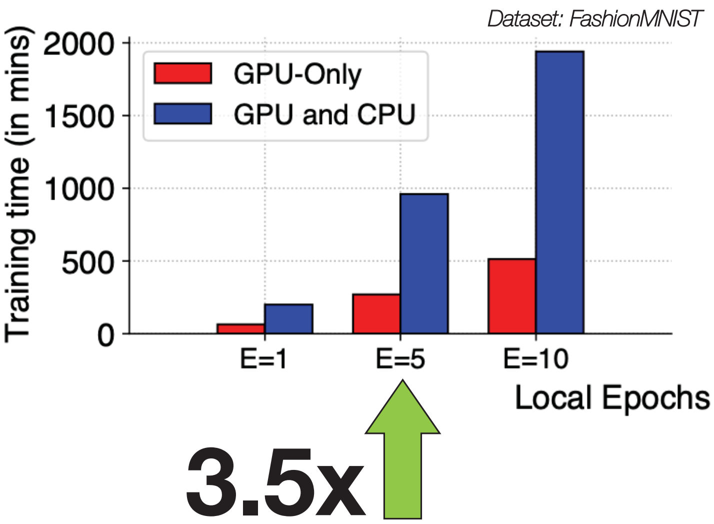


Rethinking the complete stack (and the learning algorithms)

#3 Sharing and Cooperation

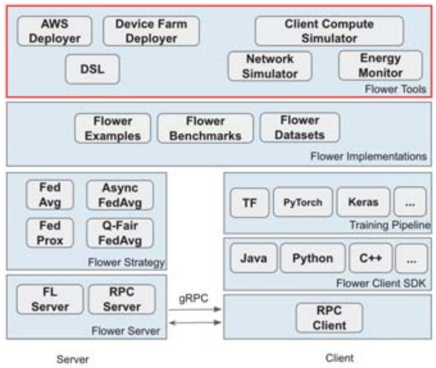
#3 Sharing and Cooperation

Federating ML Models for Mobile and Embedded Devices



Flower
http://flower.dev

BETA



68

#3 Sharing and Cooperation

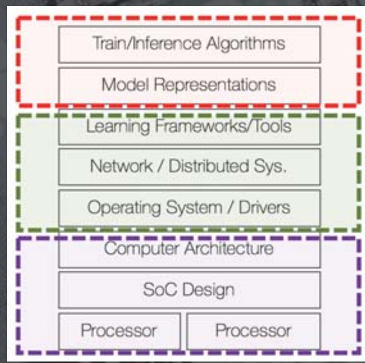
AMBITION: Maximize resource utilization through intra- and inter-device sharing and cooperation of learning algorithms

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation



Rethinking the complete stack (and the learning algorithms)

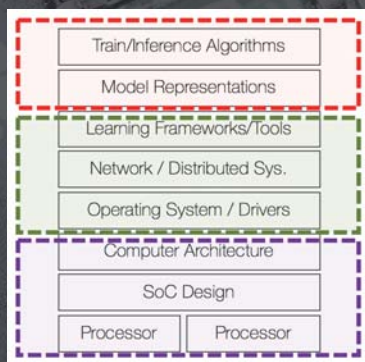
Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation

#4: Next Steps in Hardware



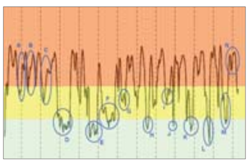
Rethinking the complete stack (and the learning algorithms)

Predictions for the Efficient ML Revolution

#1 On-Device ML goes far beyond *just* classification

#2 SOTA Accuracy will come from efficient ML Models

#3 Data Center is *replaced* by devices as the heart of ML

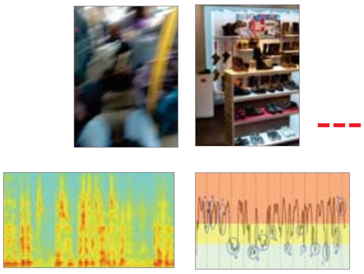


Discriminative Task

{ hand present, person sleeping }

#1 Efficient ML Prediction

On-Device ML goes far beyond classification



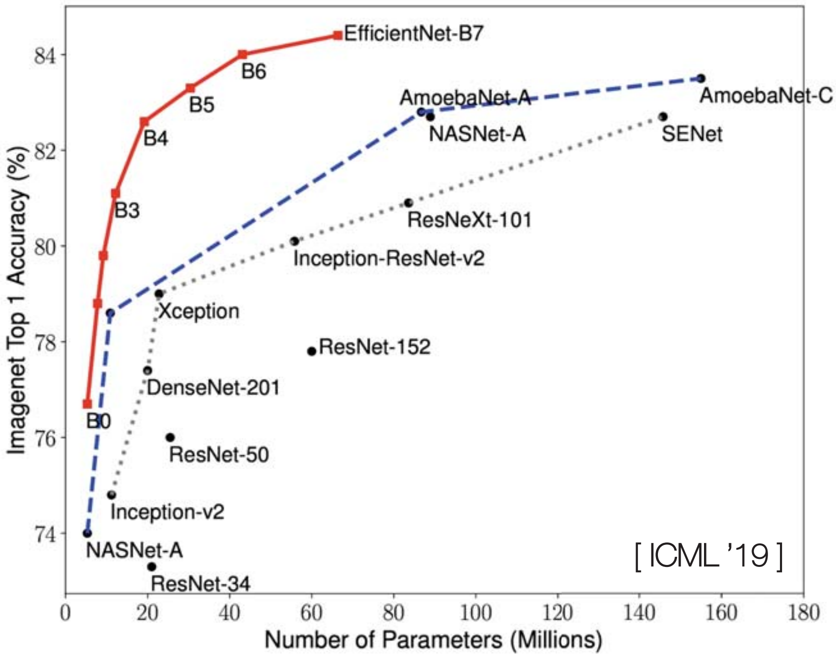
Reasoning

Common Sense

Understanding

Perception *(Discriminative)*

Cognitive
Embedded
Stack

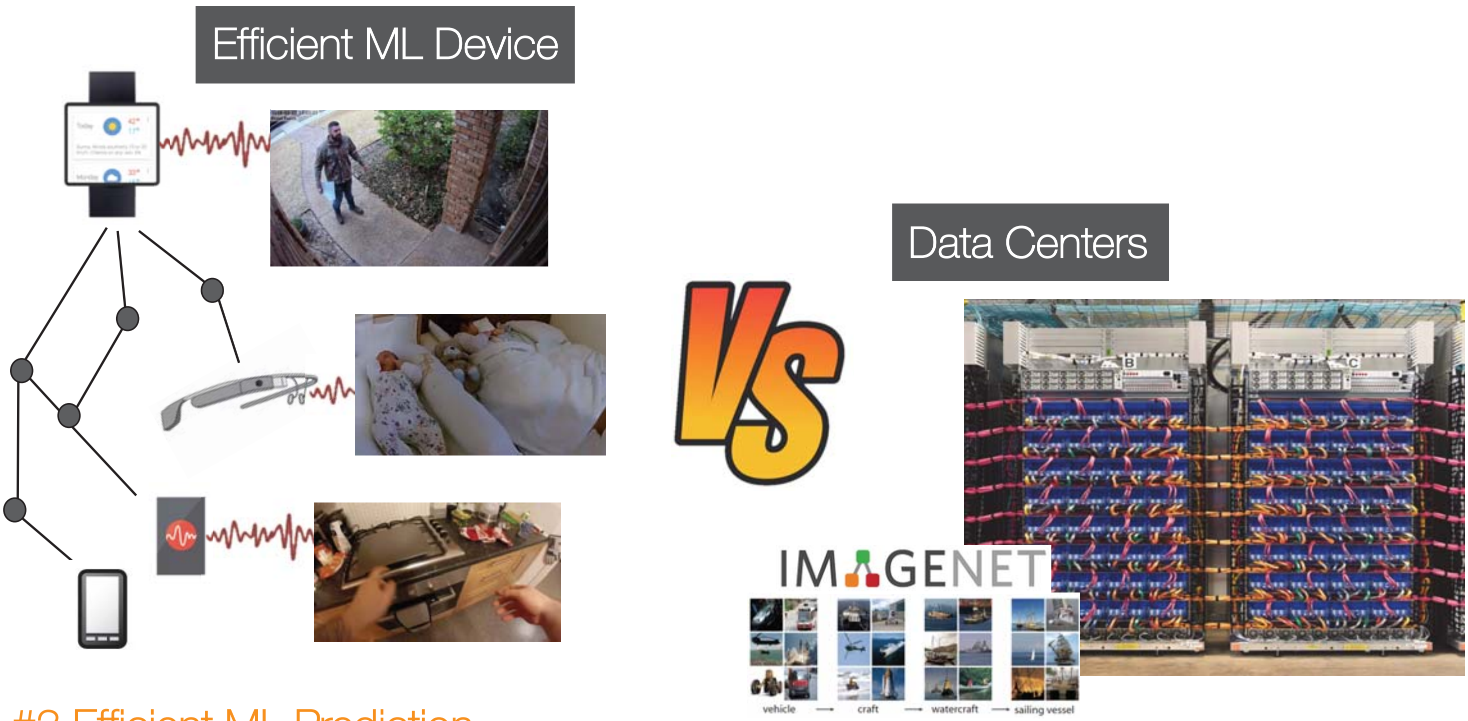


Peace Dividends of Efficiency

- Faster exploration & experimentation
- "Intractable approaches" become possible…
- Consuming more data – both *labeled and unlabeled varieties*
- Larger and larger architectures
- Able to heavily exploit automated methods like architecture search
- Brand new ML tasks

#2 Efficient ML Prediction

SOTA Accuracy will come from Efficient ML



#3 Efficient ML Prediction

Devices replace Data Centers as core of ML

UNIVERSITY OF CAMBRIDGE
Cambridge Machine Learning Systems Lab
`http://mlsys.cst.cam.ac.uk`

SAMSUNG
Samsung AI
Cambridge, UK

Select publications *(and submissions)* referenced in this talk.....

- "Flower: A Friendly Federated Learning Research Framework" – *available on Arxiv*
- "BRP-NAS: Prediction-based NAS using GCNs" – NeurIPS 2020
- "unzipFPGA: Enhancing FPGA-based CNN Engines with On-the-Fly Weight Generation" – *submitted* FPL 2020
- "Best of Both Worlds: AutoML Codesign of a CNN and its Hardware Accelerator" – DAC 2020
- "Journey Towards Tiny Perceptual Super-Resolution" – EECV 2020
- "Iterative Compression of End-to-End ASR Model using Reinforcement Learning" – ICASSP 2020
- "ShrinkML: End-to-End ASR Model Compression Using Reinforcement Learning" – INTERSPEECH 2019
- "An Empirical study of Binary Neural Networks' Optimisation" – ICLR 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
- "Squeezing Deep Learning into Mobile and Embedded Devices" – IEEE Pervasive Magazine, 16 (3), 2017
- "DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models using Wearable Commodity Hardware" – MobiSys 2017
- "Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations" – IMWUT / UbiComp 2017
- "From Smart to Deep: Robust Activity Recognition on Smartwatches using Deep Learning" – WristSense 2016
- "Sparsifying Deep Learning Layers for Constrained Resource Inference on Wearables" – SenSys 2016
- "Deepx: A software accelerator for low-power deep learning inference on mobile devices" – IPSN 2016
- "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`
`ndl32@cam.ac.uk`