Adapting AI to Available Resource in Mobile/Embedded Devices

Geoff Merrett

Implementing AI: Running AI at the Edge

12 June 2020 | KTN & eFutures Online Webinar

WHY AI AT THE EDGE?

Data Privacy

- Increased privacy if data never leaves the edge

  “ *Sending data to a central location consumes energy. Once there, **the temptation is great to keep crunching them*** [1]

Network Latency/Bandwidth/Connectivity

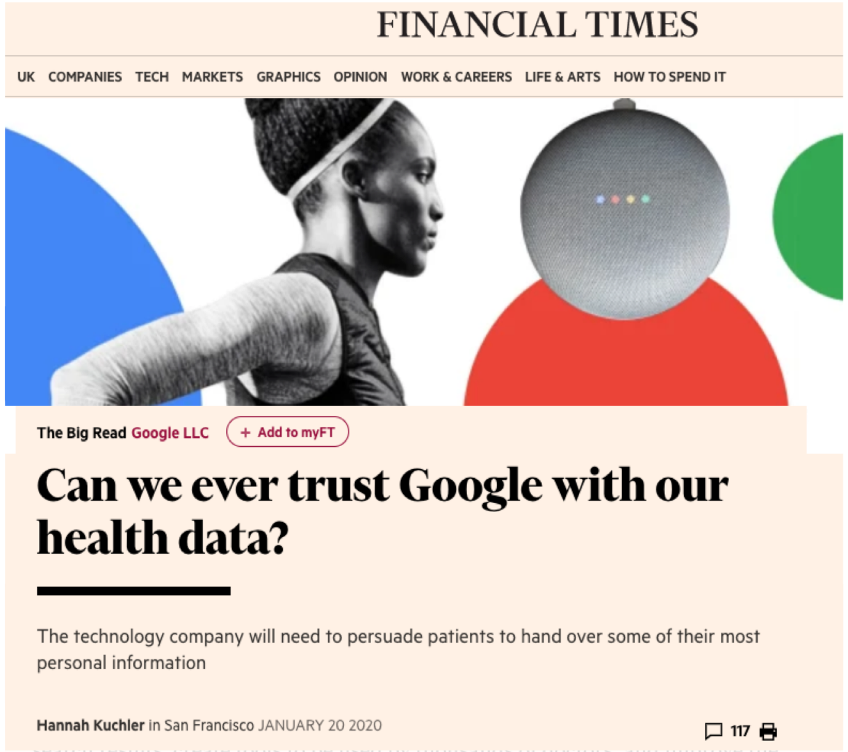
- Cloud AI *requires* good networking

  “ *Self-driving cars need very fast-reacting connections and cannot risk being disconnected; **computing needs to happen in the car itself*** [1]

  “ *Traffic lights in Las Vegas generate **60 terabytes a day** (10% of the amount Facebook collects in a day)* [1] ”

- (the edge must fulfil requirements instead though!)

[1] https://www.economist.com/special-report/2020/02/20/should-data-be-crunched-at-the-centre-or-at-the-edge


FINANCIAL TIMES

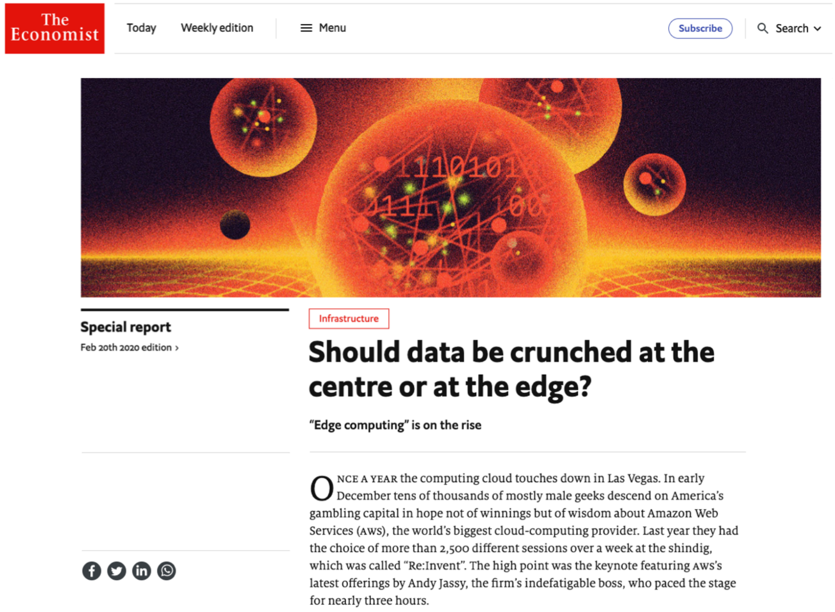
UK COMPANIES TECH MARKETS GRAPHICS OPINION WORK & CAREERS LIFE & ARTS HOW TO SPEND IT

The Big Read Google LLC + Add to myFT

Can we ever trust Google with our health data?

The technology company will need to persuade patients to hand over some of their most personal information

Hannah Kuchler in San Francisco JANUARY 20 2020 117


The Economist Today Weekly edition Menu Subscribe Search

Special report
Feb 20th 2020 edition ›

Infrastructure

Should data be crunched at the centre or at the edge?

“Edge computing” is on the rise

ONCE A YEAR the computing cloud touches down in Las Vegas. In early December tens of thousands of mostly male geeks descend on America’s gambling capital in hope not of winnings but of wisdom about Amazon Web Services (AWS), the world’s biggest cloud-computing provider. Last year they had the choice of more than 2,500 different sessions over a week at the shindig, which was called “Re:Invent”. The high point was the keynote featuring AWS’s latest offerings by Andy Jassy, the firm’s indefatigable boss, who paced the stage for nearly three hours.

WHY AI AT THE EDGE?

Power Consumption of AI

- Cloud AI consumes considerable natural resource.
  - “ *The carbon footprint of training a single AI is up to 284 tonnes of CO_2 equivalent – **5x the lifetime emissions of an average car***[2]
  - “ *An estimate puts the energy used to train the model at over **3x the yearly consumption** of the average American*[3]
  - “ *From the earliest days, the amount of computing power required by the technology doubled every two years. But from 2012 onwards, **the computing power required for today’s most-vaunted machine-learning systems has been doubling every 3.4 months***[3] ”
- An indirect benefit of moving computation to the edge, is that it *has to* be more efficient

[2] https://www.newscientist.com/article/2205779-creating-an-ai-can-be-five-times-worse-for-the-planet-than-a-car/
[3] https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning

THE DAILY NEWSLETTER
Sign up to our daily email newsletter

NewScientist

News Podcasts Video Technology Space Physics Health More Shop Tours Events

Creating an AI can be five times worse for the planet than a car

TECHNOLOGY 6 June 2019

By Donna Lu

Training artificial intelligence is an energy intensive process. New estimates suggest that the carbon footprint of training a single AI is as much as 284 tonnes of carbon dioxide equivalent – five times the lifetime emissions of an average car.

Emma Strubell at the University of Massachusetts Amherst in the US and colleagues have assessed the energy consumption required to train four large neural networks, a type of AI used for processing language.

Language-processing AIs underpin the algorithms that power Google Translate as well as OpenAI’s GPT-2 text generator, which can convincingly pen fake news articles when given a few lines of text.

Read more: AI’s dirty secret: Energy-guzzling machines may fuel global warming

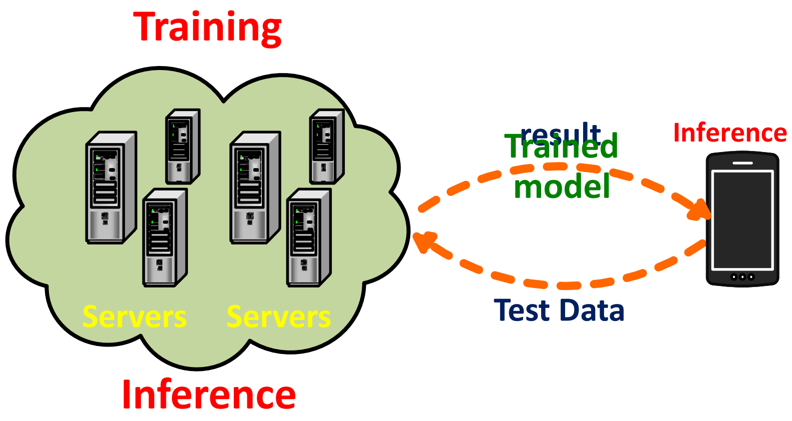
These AIs are trained via deep learning, which involves processing vasts amounts of data. “In order to learn something as complex as language, the models have to be large,” says Strubell.

A common approach involves giving an AI billions of written articles so that it learns to understands the meaning of words and how sentences are constructed.

PERFORMANCE METRICS

Inference at the Edge (/End)

- Connectivity, latency; privacy…
- *…but constrained platforms*



Platform	Computing cores	Platform-dependent metrics			Platform-independent metrics
		Execution time (ms)	Power (mW)	Energy (mJ)	Top-1 Accuracy (%)
Jetson Nano	GPU (614MHz) + A57 CPU (921MHz)	7.4	1340	9.92	71.2
	GPU (921MHz) + A57 CPU (1.43GHz)	4.93	2500	12.3	
	A57 CPU (921MHz)	69.4	878	60.9	
	A57 CPU (1.43GHz)	46.9	1490	69.9	
Odroid XU3	A15 CPU (200MHz)	1020	326	320	
	A15 CPU (1GHz)	204	846	173	
	A15 CPU (1.8GHz)	117	2120	248	
	A7 CPU (200MHz)	1780	72.4	129	
	A7 CPU (700MHz)	504	141	71.4	
	A7 CPU (1.3GHz)	280	329	92.1	

Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Optimising Resource Management for Embedded Machine Learning*. In **Design, Automation and Test in Europe Conference 2020 (DATE'20)**.

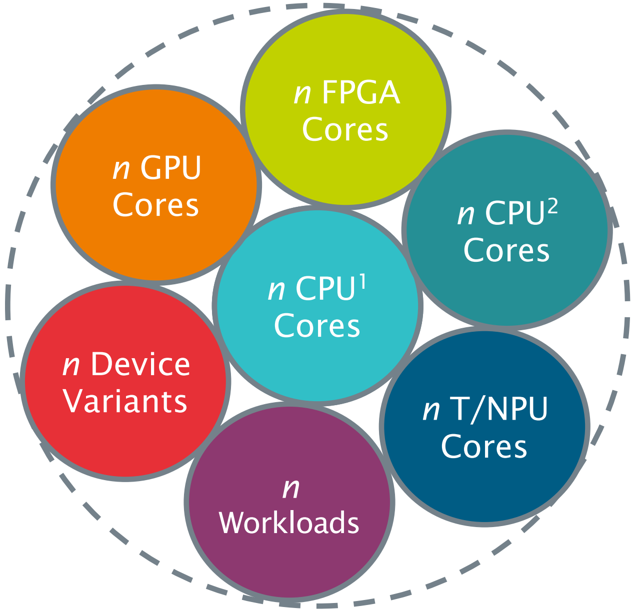
EMBEDDED AI ACCELERATION

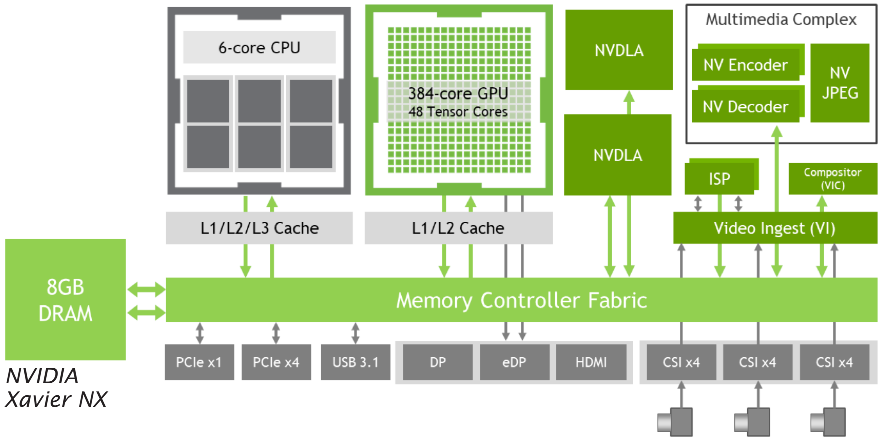
- General/specialist compute units for AI rapidly increasing
- Some mobile/embedded AI systems are reasonably static…
- …however, others aren't
  - General purpose systems
  - Multi-tenant systems
  - 'Adaptive' AI/event-driven operation
  - *etc*

UNIVERSITY OF Southampton

SYSTEM RESOURCE MANAGEMENT

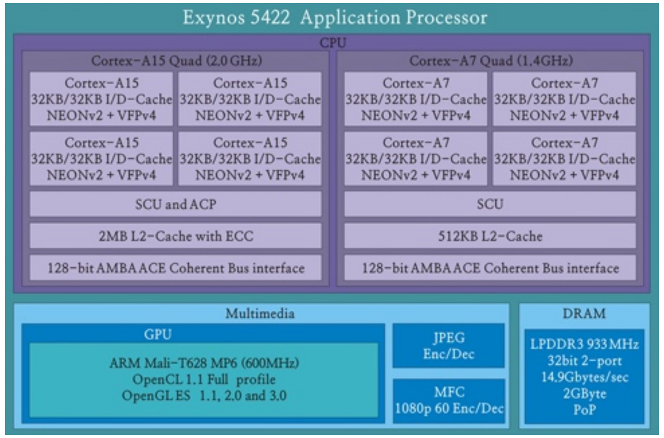
- Complexity of hardware-software interaction has grown



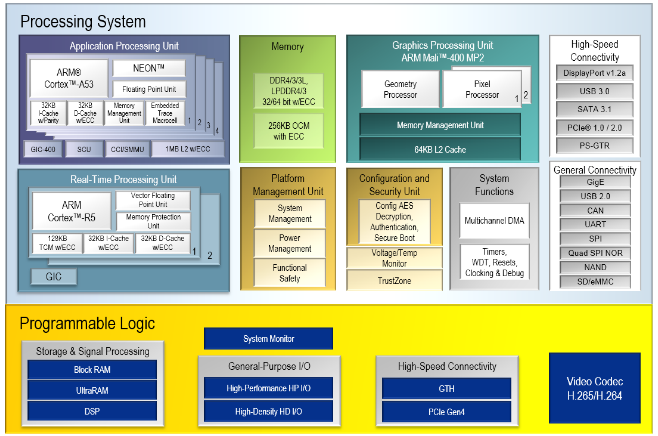


NVIDIA Xavier NX

- Managing resources is no longer trivial, yet is increasingly needed



Samsung Exynos 5422



Xilinx Zynq Ultrascale+

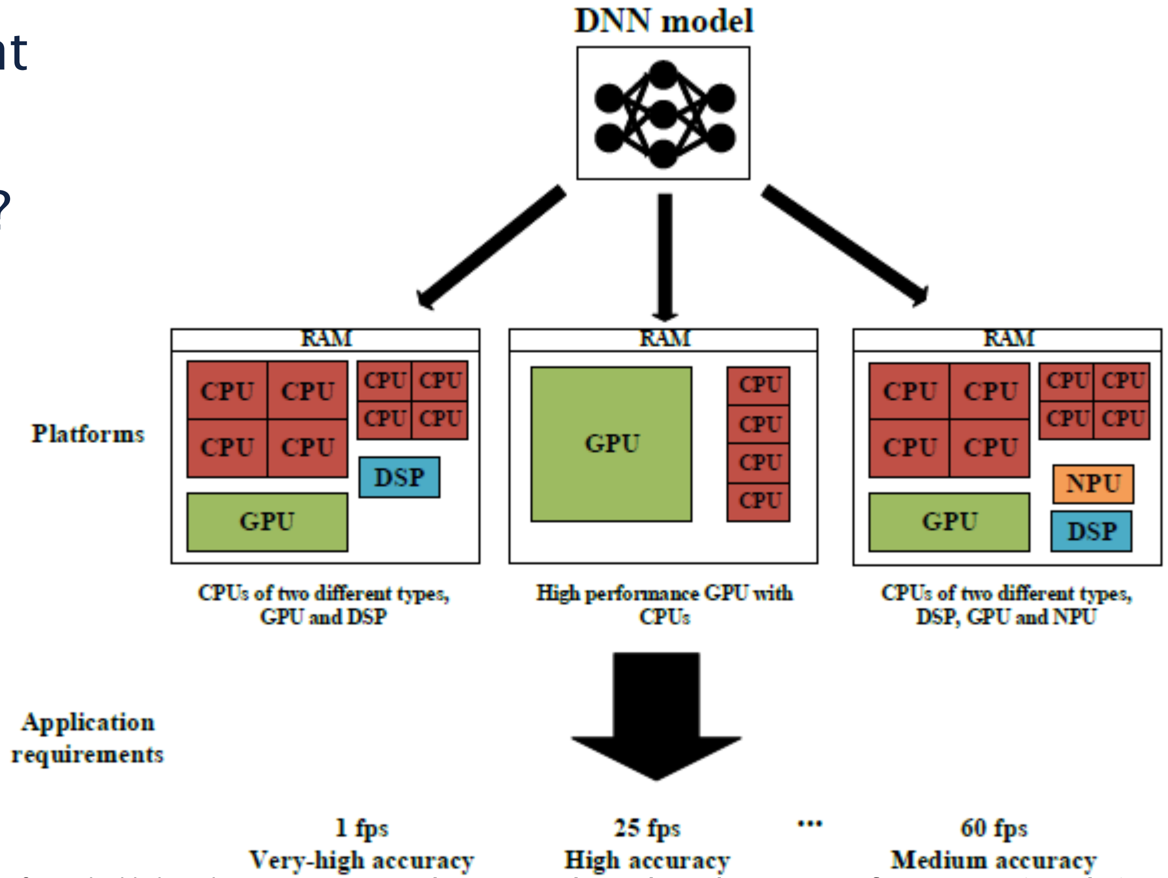
DESIGN-TIME CHALLENGES

Platform Diversity

How can we develop DNN models that can:

1. operate across a wide range of different heterogeneous platforms, and
2. meet diverse application requirements?

- Existing design-time approaches such as static model pruning compress the model to approximately the ‘right size’.



Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Optimising Resource Management for Embedded Machine Learning*. In **Design, Automation and Test in Europe Conference 2020 (DATE'20)**.

RUN-TIME CHALLENGES

Workload Diversity

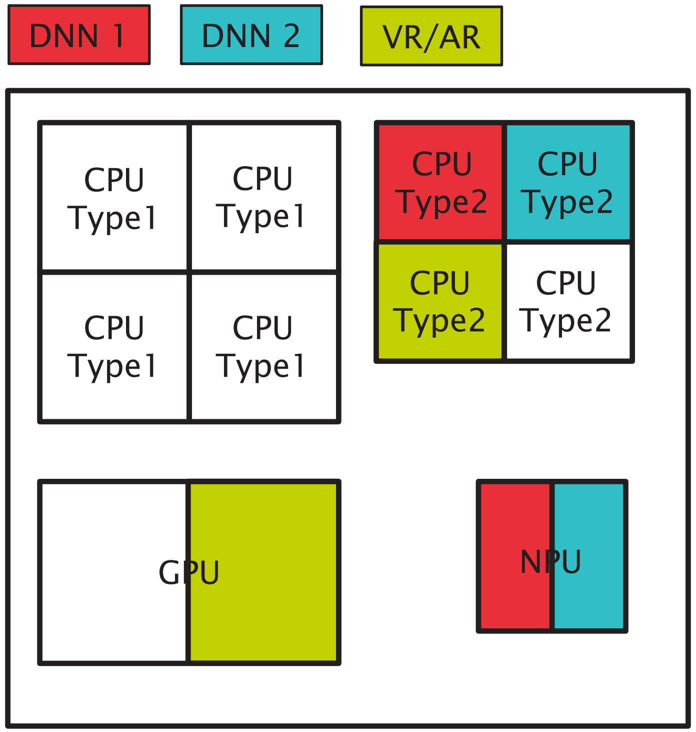
How can we perform inference while:

1. meeting timing requirements?
2. meeting power/energy requirements?
3. meeting accuracy requirements?

How can we do this:

- while executing another DNN model at the same time?
- while executing other foreground/ background tasks at the same time?

We need dynamic DNNs…



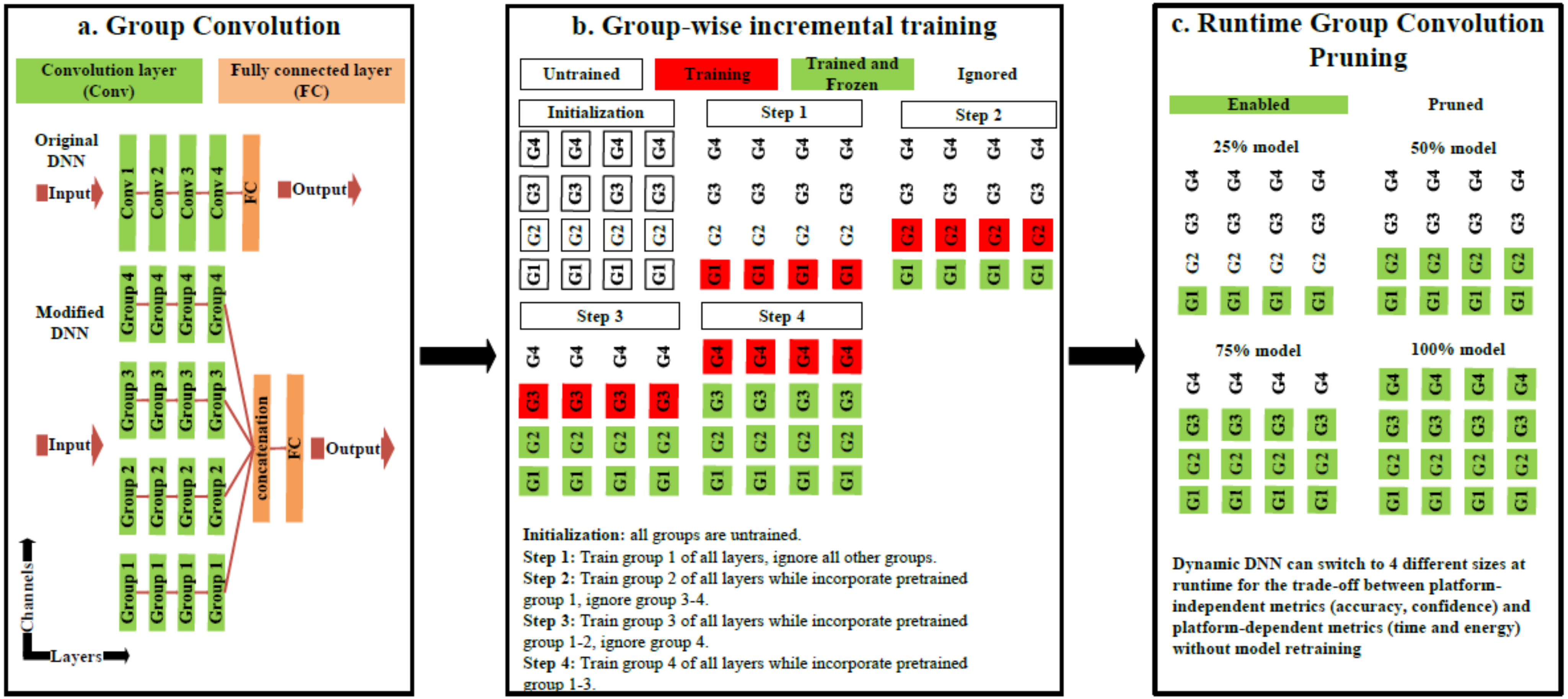
Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Optimising Resource Management for Embedded Machine Learning.* In **Design, Automation and Test in Europe Conference 2020 (DATE'20)**.

UNIVERSITY OF Southampton

DYNAMIC DNNs

Incremental Training with Group Convolution Pruning

L. Xun *et al*. *Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms*. In **Workshop on Machine Learning for CAD (MLCAD'19)**.



DYNAMIC DNNs

Experimental Setup

Model: Modified AlexNet (~320kB)

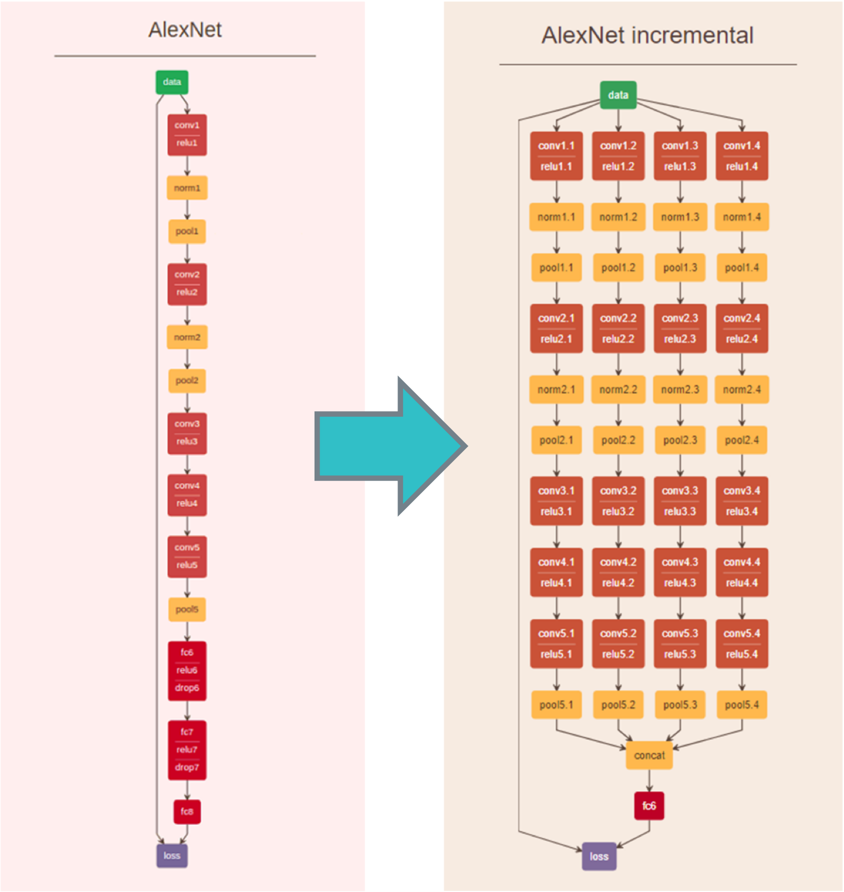
Dataset: CIFAR10

- 32*32*3 images in 10 classes
- 50,000 training and 10,000 testing images

Framework: Caffe

Hardware:

- Odroid XU3
  - CPU: 4x Arm A15 (f = 0.2–2 GHz) + 4x Arm A7 (f = 0.2–1.4 GHz)
  - GPU: Mali-T628 (not used in these experiments)
- Nvidia Jetson Nano
  - CPU: 4x Arm A57 (f = 0.9, 1.4 GHz)
  - GPU: 128x CUDA core Maxwell (f = 0.6, 0.9 GHz)



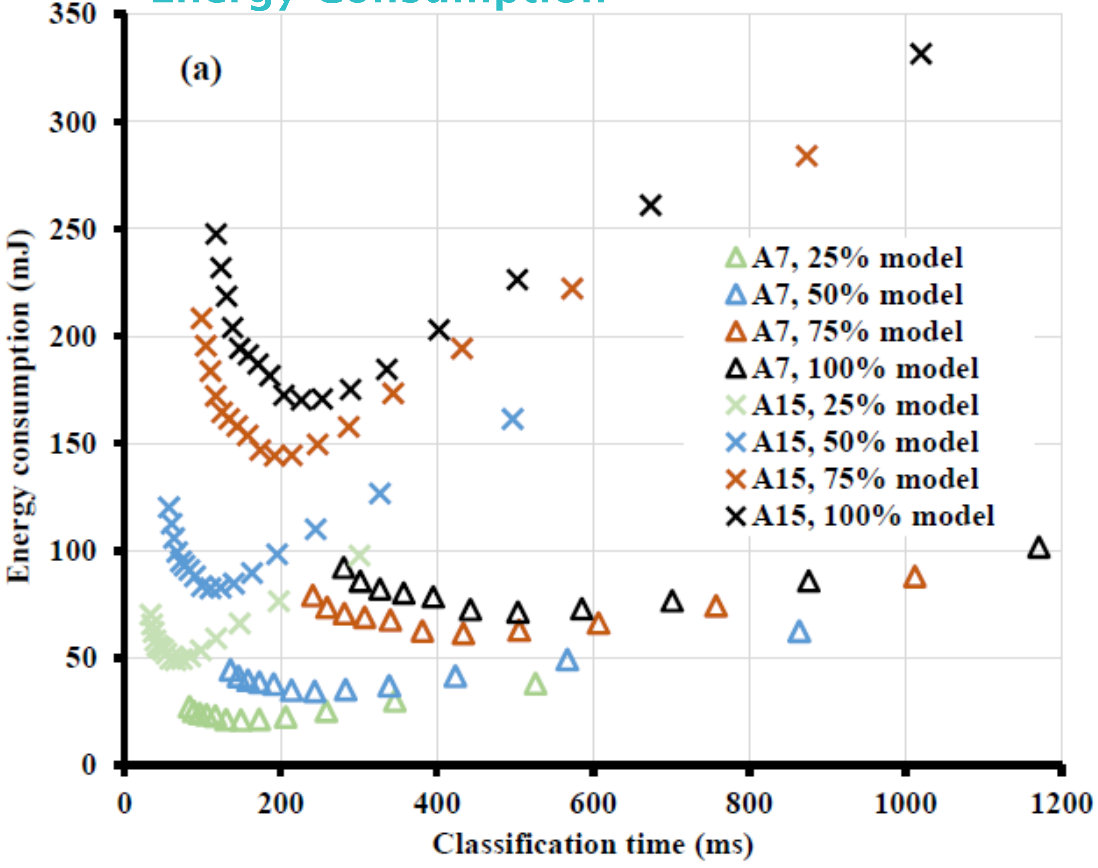
Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms.* In **Workshop on Machine Learning for CAD (MLCAD'19)**.

Southampton UNIVERSITY OF

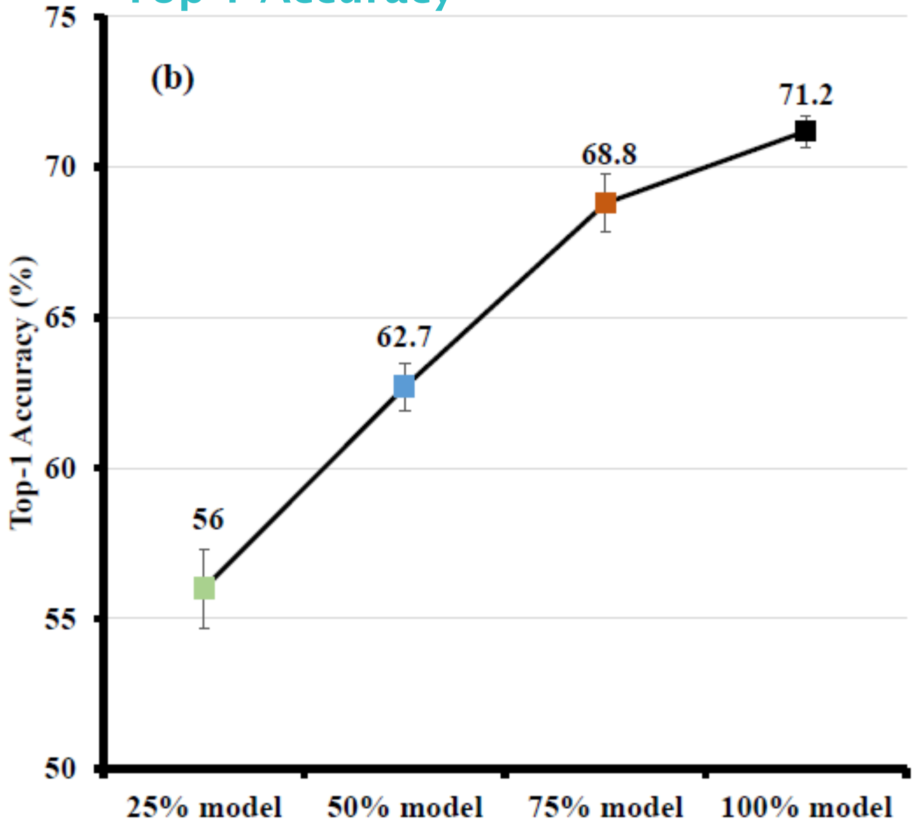
DYNAMIC DNNs

Results: DVFS and Task Mapping (Odroid XU3)

Energy Consumption



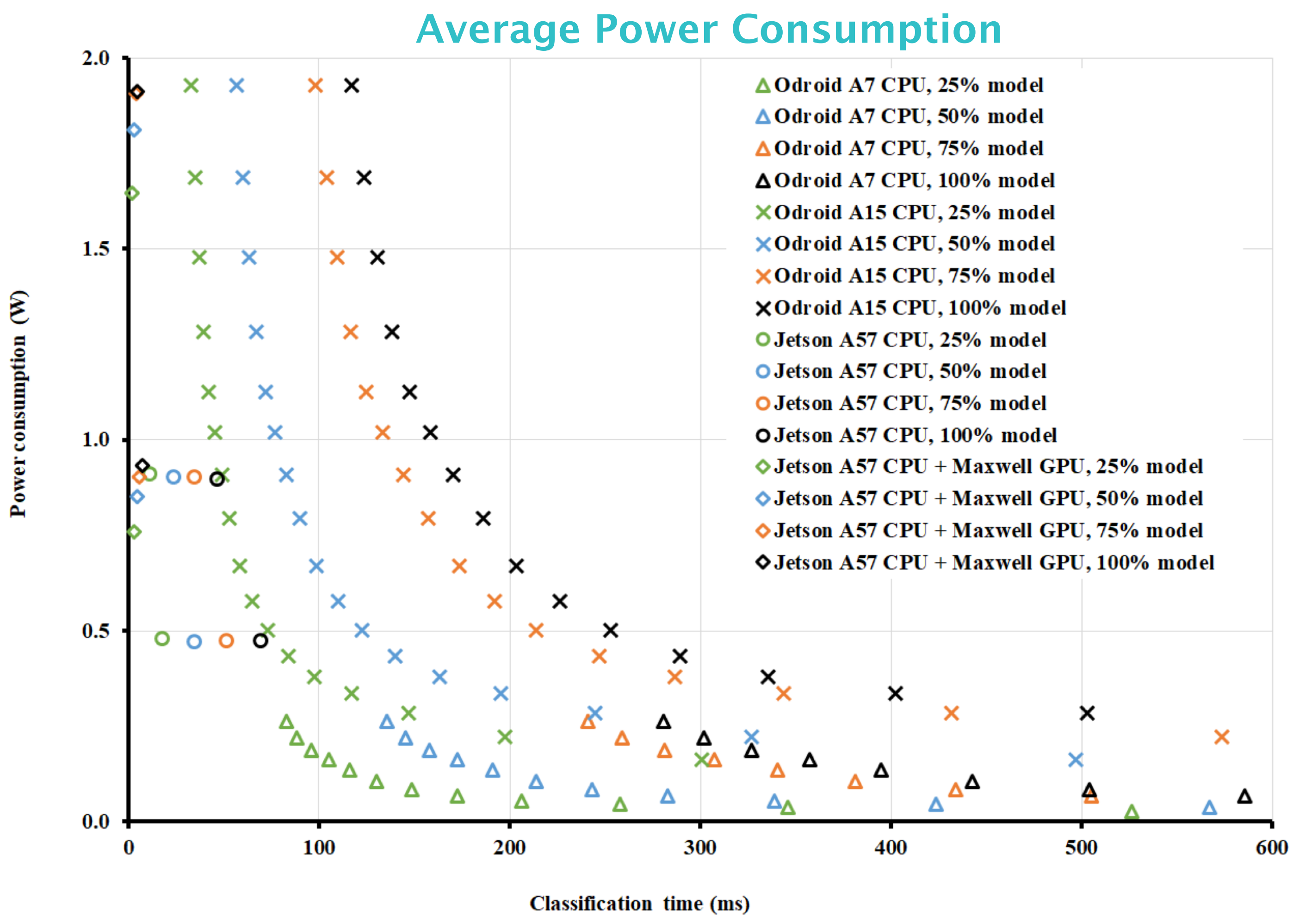
Top-1 Accuracy



Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms.* In **Workshop on Machine Learning for CAD (MLCAD'19)**.

DYNAMIC DNNs

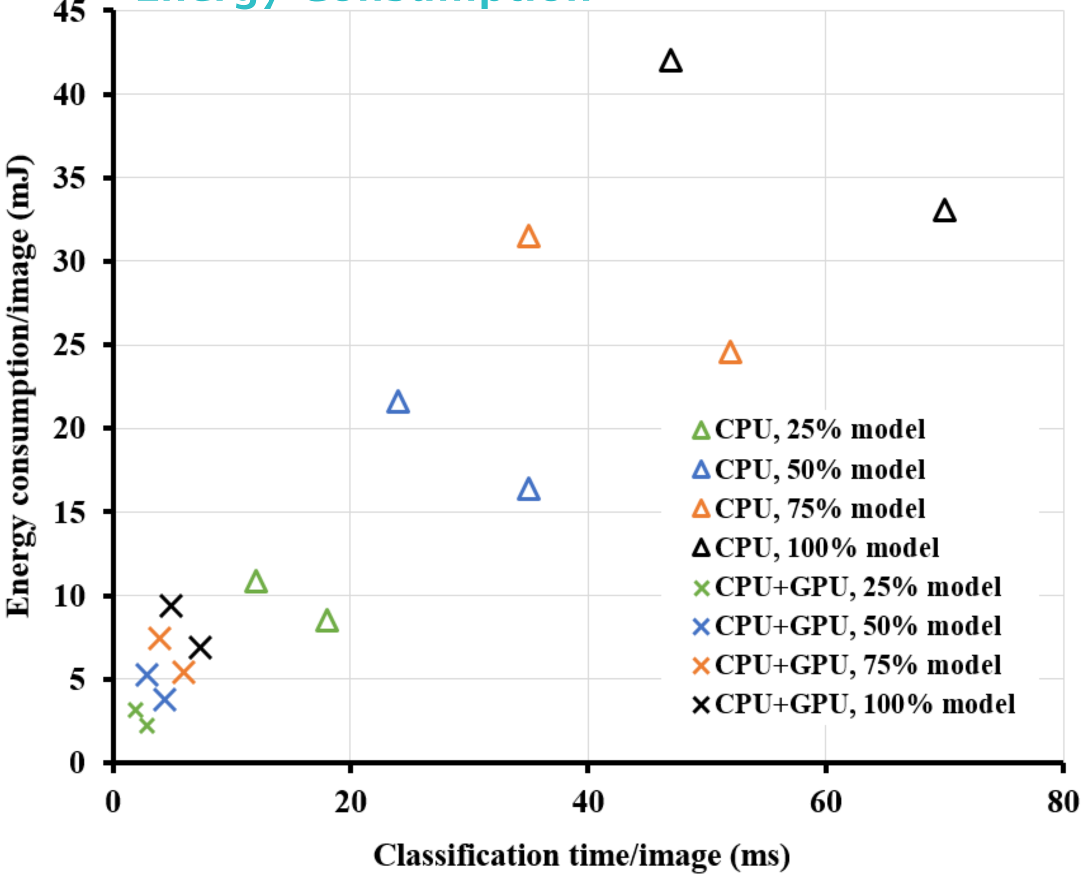
Results: Power Consumption



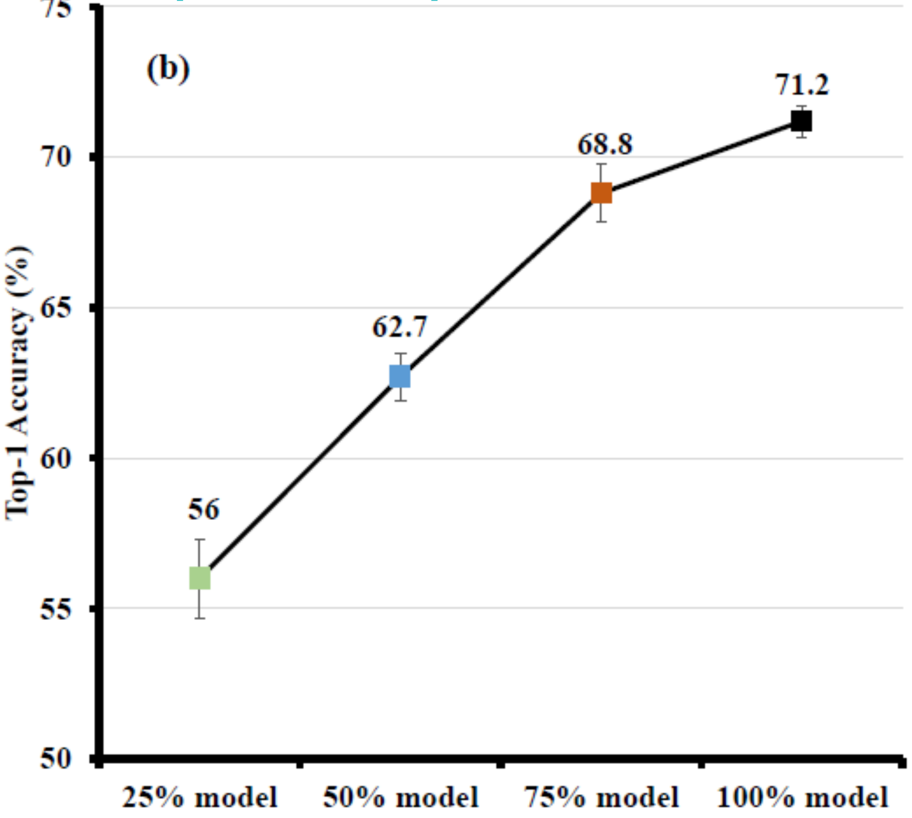
DYNAMIC DNNs

Results: DVFS and Task Mapping (Jetson Nano)

Energy Consumption



Top-1 Accuracy



Xun, Lei, Tran-Thanh, Long, Al-Hashimi, Bashir and Merrett, Geoff (2020) *Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms.* In **Workshop on Machine Learning for CAD (MLCAD'19)**.

RUNTIME POWER MANAGEMENT

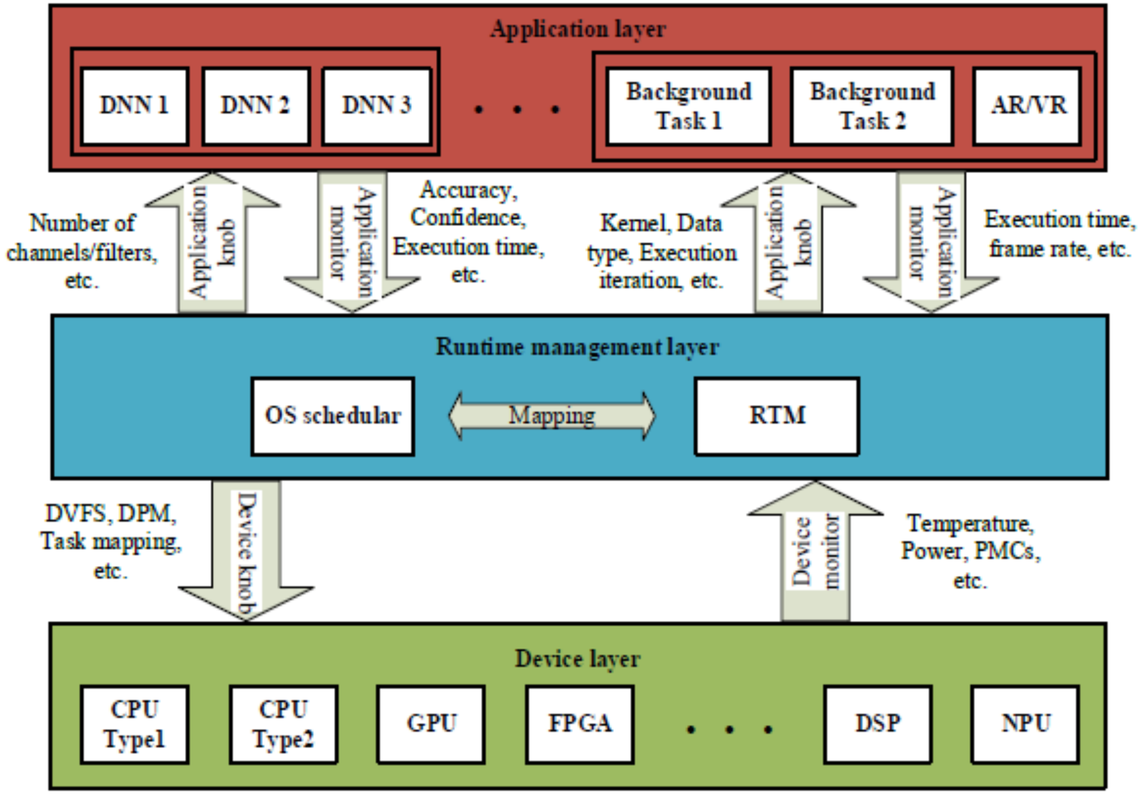
www.prime-project.org

Runtime Management (RTM)

- System software to react and predict
- Controls/'knobs'
- 'Monitors'/sensors

RTM to coordinate/balance…

- Mapping to heterogeneous PEs
- Response to environmental factors
- Power consumption/battery life
- (concurrently) Executing tasks
- Application(s) requirements
- User requirements/QoE



Bragg, Graeme McLachlan, Leech, Charles R., Balsamo, Domenico, Davis, James J., Weber Wachter, Eduardo, Merrett, Geoff, Constantinides, George A. and Al-Hashimi, Bashir (2018) *An application- and platform-agnostic control and monitoring framework for multicore systems.* **3rd International Conference on Pervasive and Embedded Computing**, Portugal. 29 - 30 Jul 2018.

Photo by Patrick Schneider on Unsplash

CONCLUSIONS

- AI is moving to the edge…

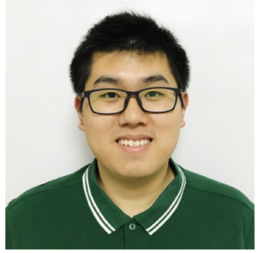
> “*If machine learning is going to be deployed at a global scale, most of the computation will have to be done in users’ hands, ie in their smartphones* [3]

- …but available resources on edge platforms are typically both constrained and time-varying
- We need improved approaches to manage resources in systems while providing *acceptable* performance

> “*Companies will learn to make trade-offs between accuracy and computational efficiency, though that will have unintended, and antisocial, consequences too* [3] ”

[3] https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning

ACKNOWLEDGEMENTS



Lei Xun (PhD student)
w https://www.ecs.soton.ac.uk/people/lx2u16
@XunLei_CHN

International Centre for Spatial Computational Learning (EPSRC)
w https://spatialml.net/
@spatialmlnet

Power and Reliability in Many-Core Embedded Systems (EPSRC)
w https://www.prime-project.org
@prime_programme

YOUR QUESTIONS

Professor Geoff Merrett
Head of Centre for IoT and Pervasive Systems

e: gvm@ecs.soton.ac.uk
w: www.geoffmerrett.co.uk
@g_merrett

UNIVERSITY OF
Southampton