Machine Learning on Embedded Things

"Smaller, and Easier"

Lisa Ong

Principal Lecturer and Consultant, NUS ISS

ML on Embedded Things

- Value Proposition
- Challenges
- Bridging the Gap
- Demo
- Summary

Value Proposition

Why is ML is going to Embedded Things

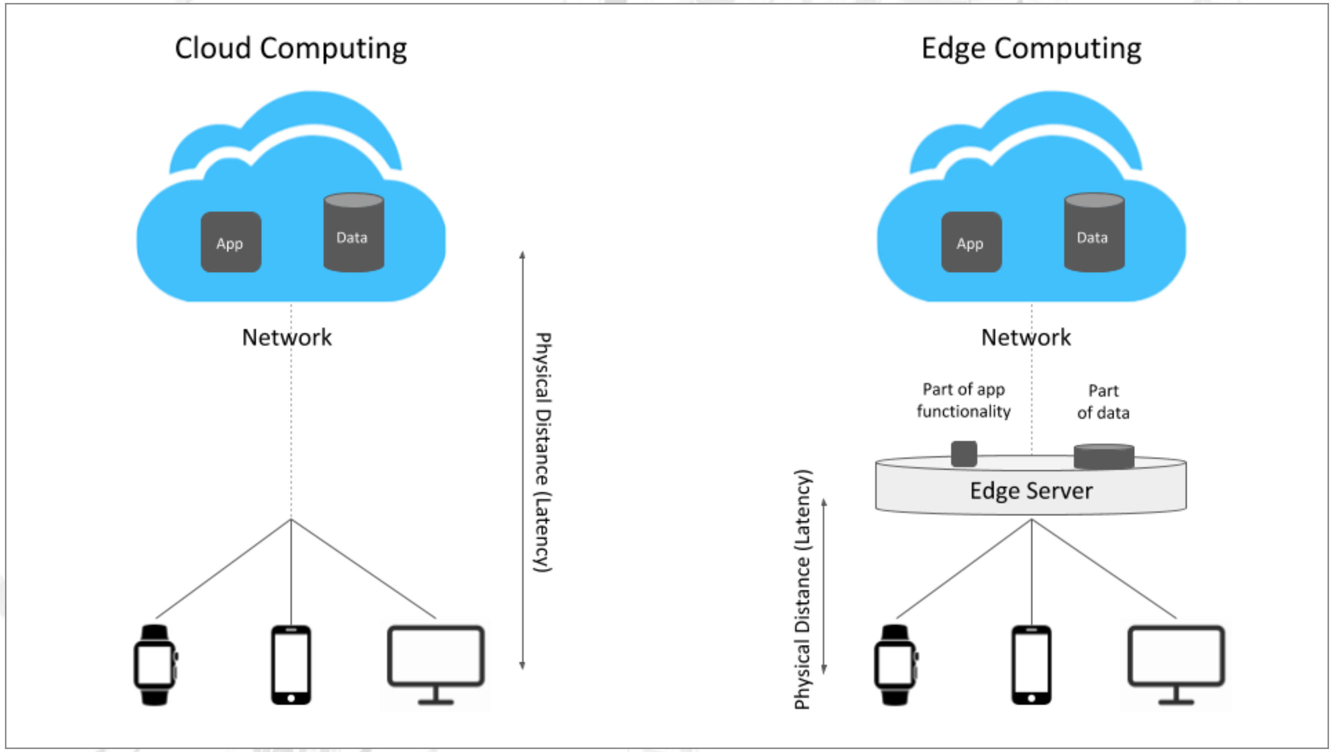
Edge Computing is the New Cloud (1)

More Responsive

- Shifts processing nearer to the Embedded Things (sensors, actuators, controllers) layer
- Lower transmission latencies
- Local decision-making

More Resilient

Increased Privacy



trantorinc.com

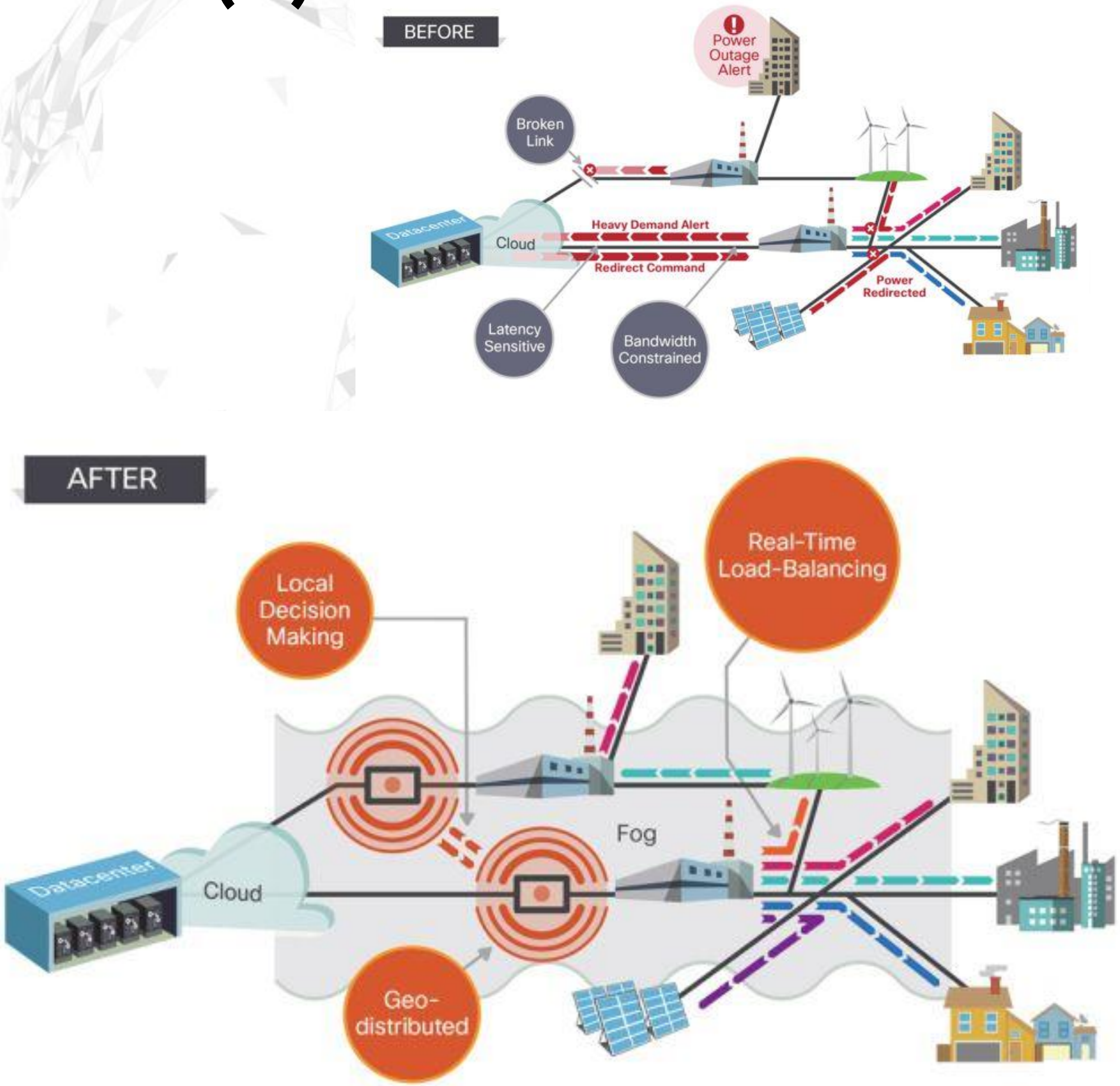
Edge Computing is the New Cloud (2)

More Responsive

More Resilient

- Divide-n-conquer: processing functions shared across multiple nodes and layers
- Fewer bottlenecks, failure points
- Scalable: vertical and horizontally

Increased Privacy



sites.ndtv.com

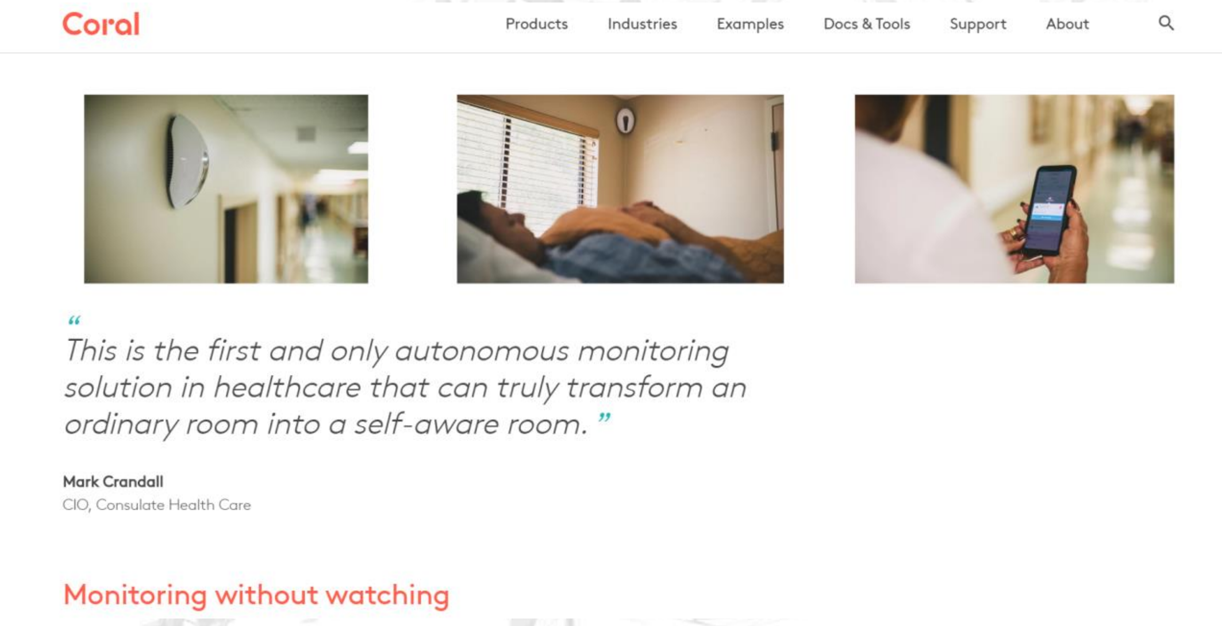
Edge Computing is the New Cloud (3)

More Responsive

More Resilient

Increased Privacy

- Sensitive-data is processed locally
- Only transmit the result to the Cloud
- Access and scope can be limited using on physical boundaries
- Only devices within range will receive the data



care.ai, coral.ai

Edge Computing Market Value

US$3.5B	Global Value in 2019 (Grand View Research)
US$43.4B	Projected Global Value by 2027 @ 37.4% growth p.a. Grand View Research)
29B	Connected devices by 2022 (Telecommunications Industry Association)
75%	Enterprise Data created and processed at the Edge by 2025. Compared to 10% in 2018 (Gartner)
30ms	Latency of early 5G networks (Verizon)
10-20ms	Projected latency of future 5G networks (IBM)

enterprisersproject.com

Edge Computing and ML

Cloud Computing uses AI, specifically Deep Learning to solve complex problems using Big Data

Volume of Edge data is exponentially bigger because of the growing number of connected devices

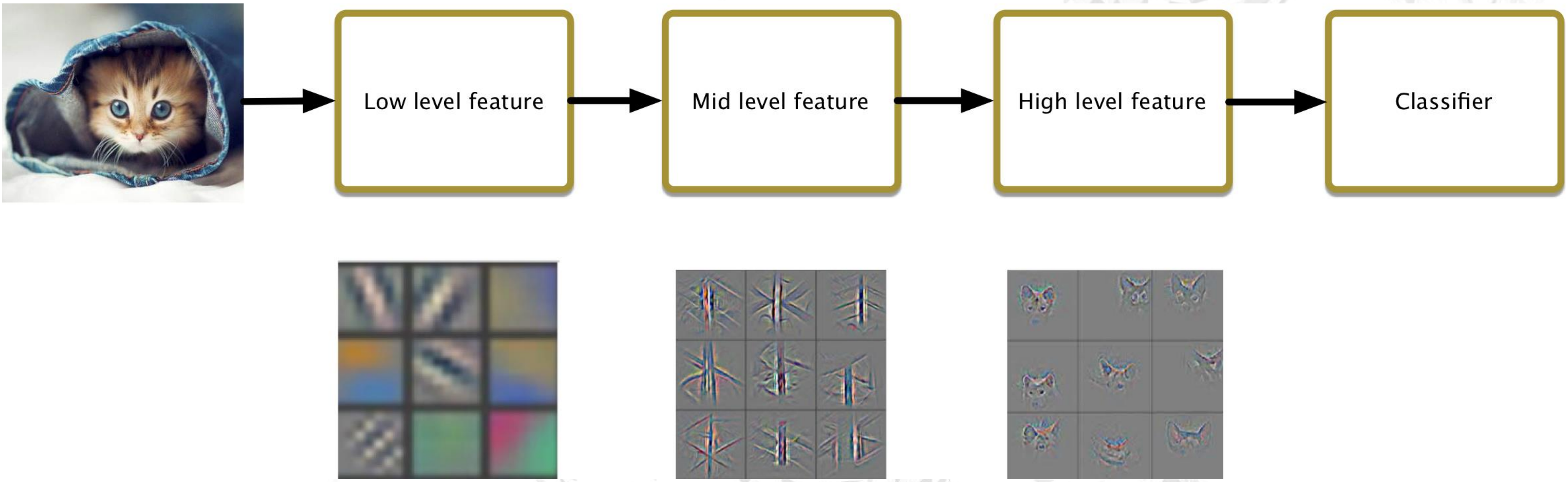
Can we “port” Deep Learning to run on Edge Computing?

Challenges

Poll: What is so difficult about running ML on the Edge?

Deep Neural Networks are Heavyweight (1)

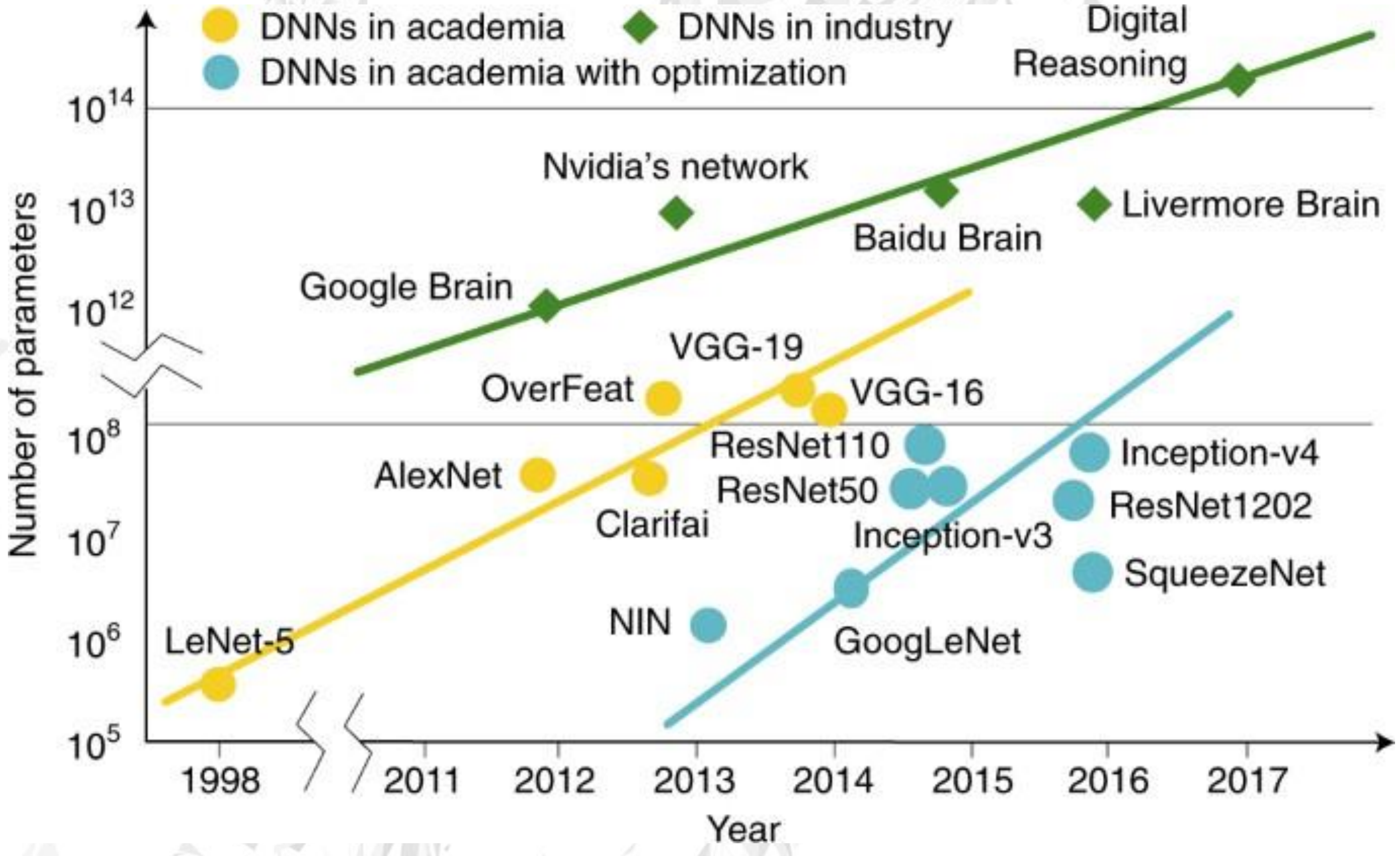
Deep Neural Networks perform successive layers of mathematical operations on the input data to get a result



arxiv.org/abs/2001.03460

Deep Neural Networks are Heavyweight (2)

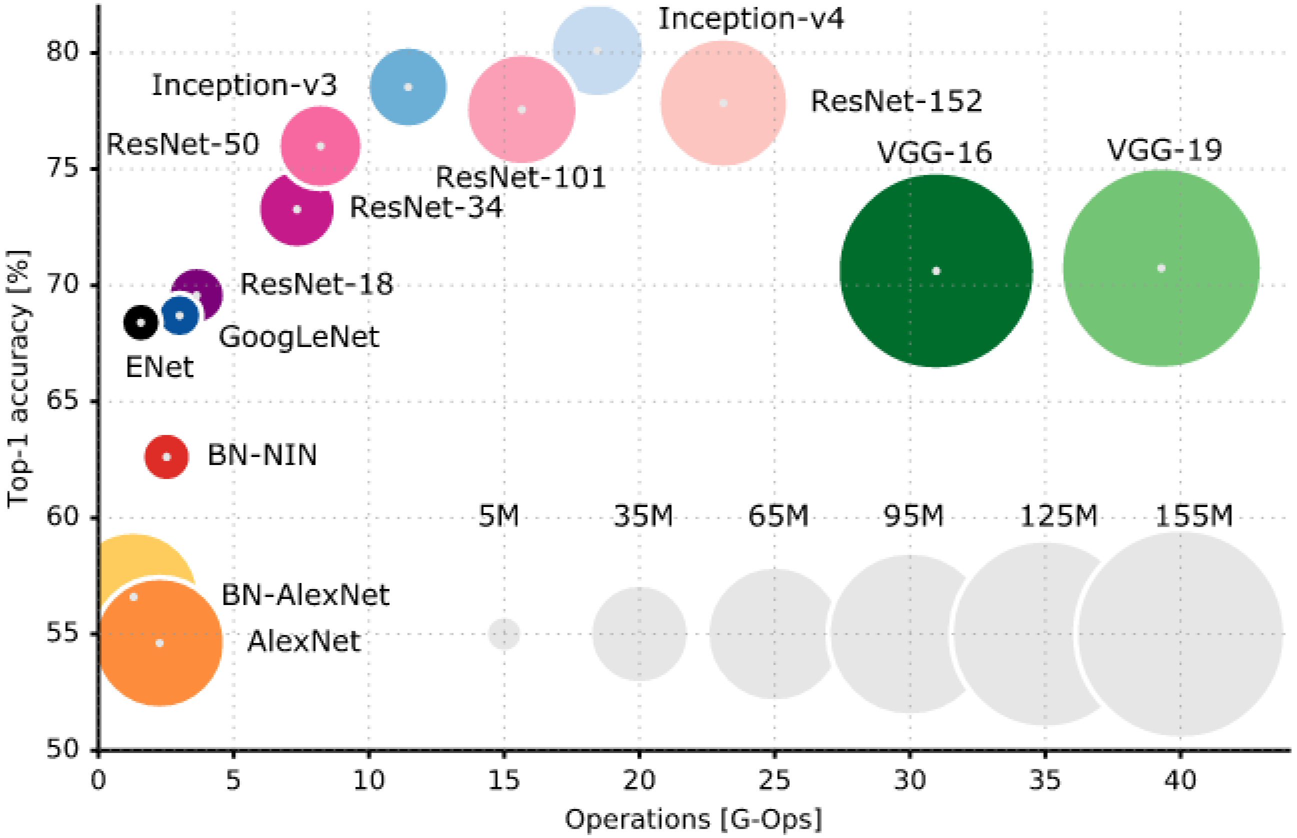
- Each layer can contain many parameters
- More complex domains can require more parameters
  - Recognize Cats vs Recognize Human Behavior
- More parameters = larger neural networks
- 10^6 float32 params ~ 30MB



nature.com

Deep Neural Networks are Heavyweight (2)

- Large neural networks also involve more operations
  - Measured in Floating Point Operations (FLOPs)
- Each parameter may be used multiple times
  - FLOPs is proportional to the size of the network



kdnuggets.com

Deep Neural Networks use Heavyweight Hardware

Hardware Comparison

The table below shows the key hardware differences between Nvidia's P100 and V100 GPUs.

Processor	SMs	CUDA Cores	Tensor Cores	Frequency	TFLOPs (double)[1]	TFLOPs (single)[1]	TFLOPs (half/Tensor)[1,2]	Cache	Max. Memory	Memory B/W
Nvidia P100 PCIe (Pascal)	56	3,584	N/A	1,126 MHz	4.7	9.3	18.7	4 MB L2	16 GB	720 GB/s
Nvidia V100 PCIe (Volta)	80	5,120	640	1.53 GHz	7	14	112	6 MB L2	16 GB	900 GB/s

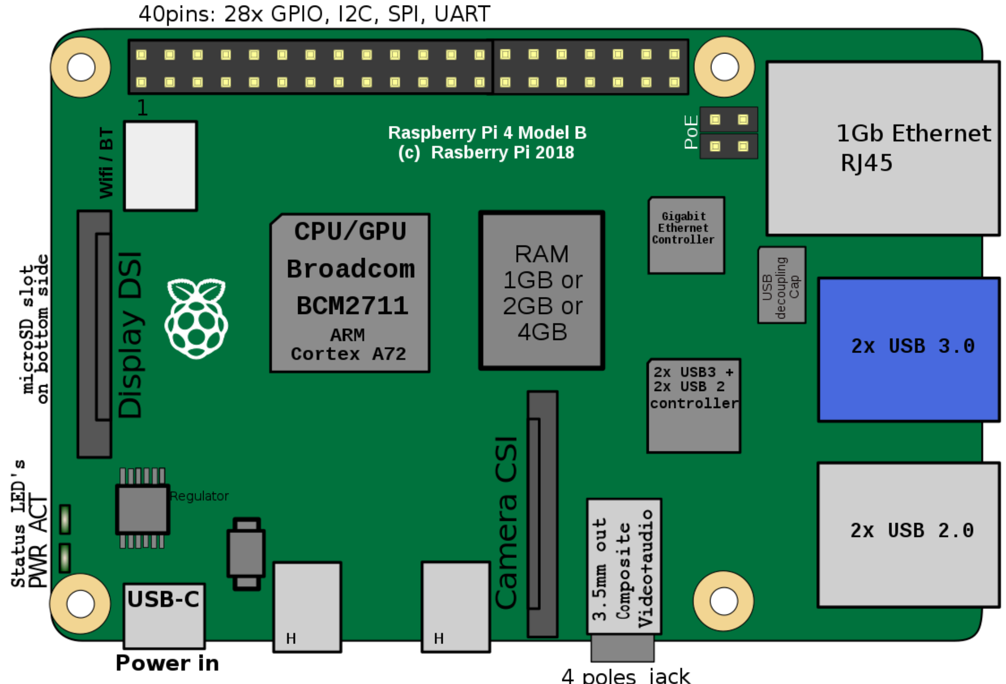
[1]Note that the FLOPs are calculated by assuming purely fused multiply-add (FMA) instructions and counting those as 2 operations (even though they map to just a single processor instruction).

[2]On P100, half-precision (FP16) FLOPs are reported. On V100, tensor FLOPs are reported, which run on the Tensor Cores in mixed precision: a matrix multiplication in FP16 and accumulation in FP32 precision.

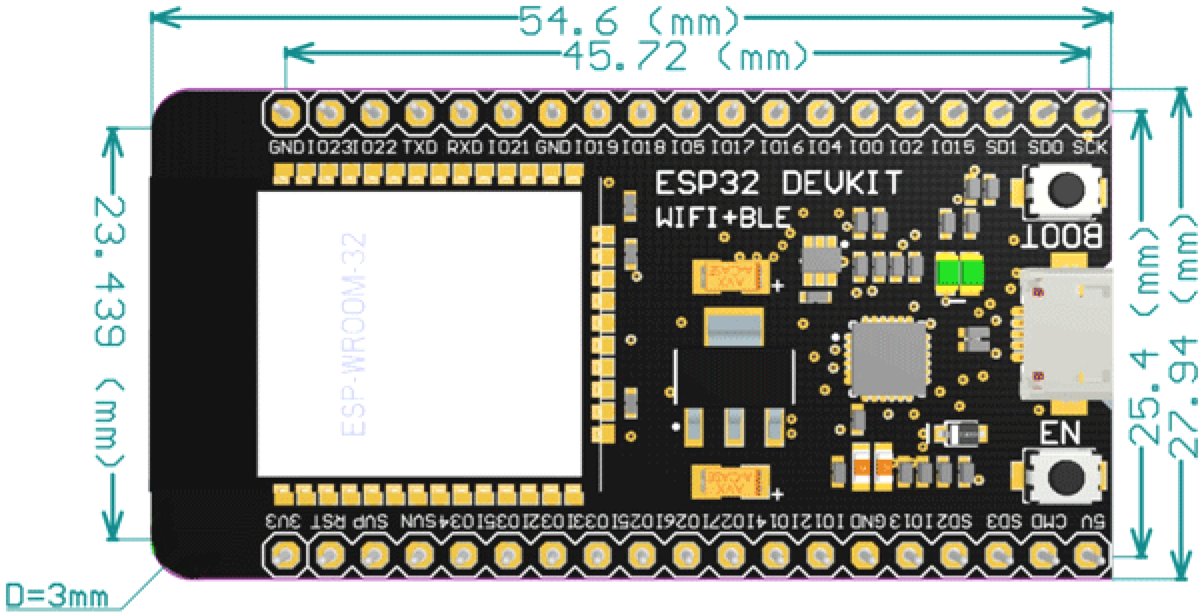
1 Tera FLOP = 1 Trillion FLOPs

xcelerit.com

Embedded IoT Devices are Constrained Hardware



wikipedia



components101.com

Raspberry Pi 4 Model B

4 x Cortex A72 @ 1.5GHz

8 FLOPs (FP16)

1-8GB RAM

ESP32-WROOM-32

40 MHz

4MB SPI flash, 320kB DRAM

#ISSLearningFest

STM32

STM32F4 MCU Series
32-bit Arm® Cortex®-M4 – Up to 180 MHz

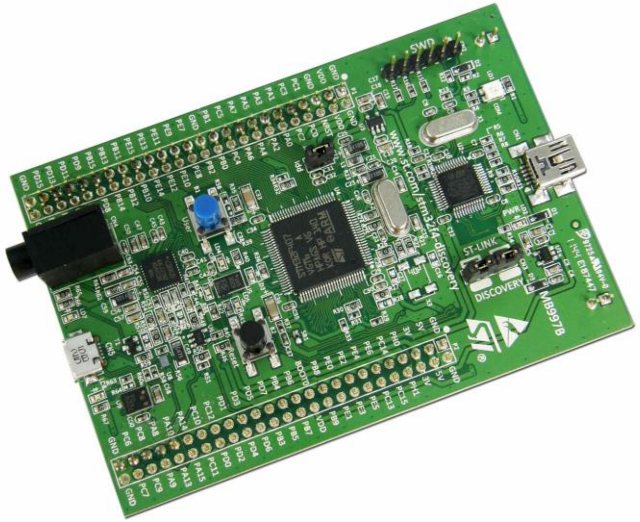
LONGEVITY 10 YEARS COMMITMENT

Product lines	F_{CPU} (MHz)	Flash (Kbytes)	RAM (KB)	Ethernet I/F IEEE 1588	2x CAN	Camera I/F	SDRAM I/F	Dual Quad-SPI	SAI	SPDIF RX	Chrom-ART Graphic Accelerator™	TFT LCD Controller	MIPI DSI
Advanced lines													
STM32F469[2]	180	512 K to 2056 K	384	•	•	•	•	•	•		•	•	•
STM32F429[2]	180	512 K to 2056 K	256	•	•	•	•		•		•	•	
STM32F427[2]	180	1024 K to 2056 K	256	•	•	•	•		•		•		
Foundation lines													
STM32F446	180	256 K to 512 K	128		•	•	•	•	•	•			
STM32F407[2]	168	512 K to 1024 K	192	•	•	•							
STM32F405[2]	168	512 K to 1024 K	192		•								

Product lines	F_{CPU} (MHz)	Flash (Kbytes)	RAM (KB)	RUN current (µA/MHz)	STOP current (µA)	Small package (mm)	FSMC (NOR/PSRAM/LCD support	QSPI	DFSDM	DAC	TRNG	DMA Batch Aquisition Mode	SB 2.0 OTG FS

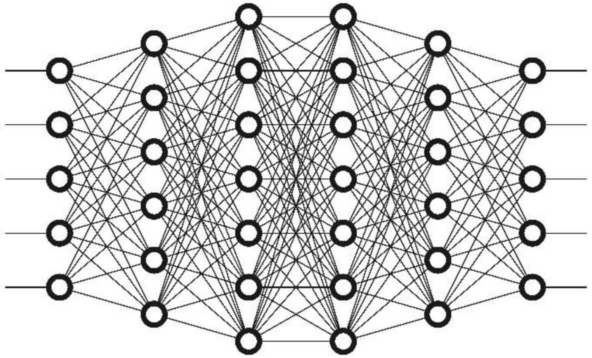
- ART Accelerator™
- SDIO
- USART, SPI, I²C
- I²S + audio PLL
- 16 and 32-bit timers
- 12-bit ADC (0.41 µs)
- True Random Number Generator
- Batch Acquisition Mode
- Low voltage 1.7 to 3.6 V
- Temperature:
- -40 °C to 125 °C

st.com



blog.tkjelectronics.dk

Cortex M4
168-180MHz
Max 384kB RAM
Max 2MB Flash

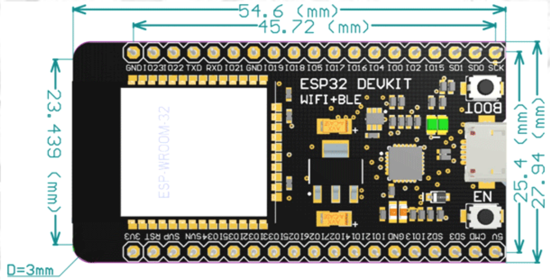


www.datanami.com

Millions of parameters,
GB sizes

Billions of FLOPs,
usually parallelized on
TFLOP processors

earlyadopter.com



components101.com

Constrained
320kB - 8GB RAM
40MHz - 4x1.5GHz

Up to 8 FLOPs (RPi4)
Limited or no parallelization

Bridging the Performance Gap

Making it easier to run DNNs on Embedded Things

Bridging the Gap (1)

Shrink Deep Neural Networks

- TensorFlow Lite quantization
  - Float32 to Int8
  - 4x smaller, 3x+ speedup
  - Mostly post-training
- “Once-for-all” network
  - Train specialized architectures concurrently
  - Strive to optimize all at the same time

Power-up Embedded Things

Optimise Tools & Workflows

Representation for quantized tensors

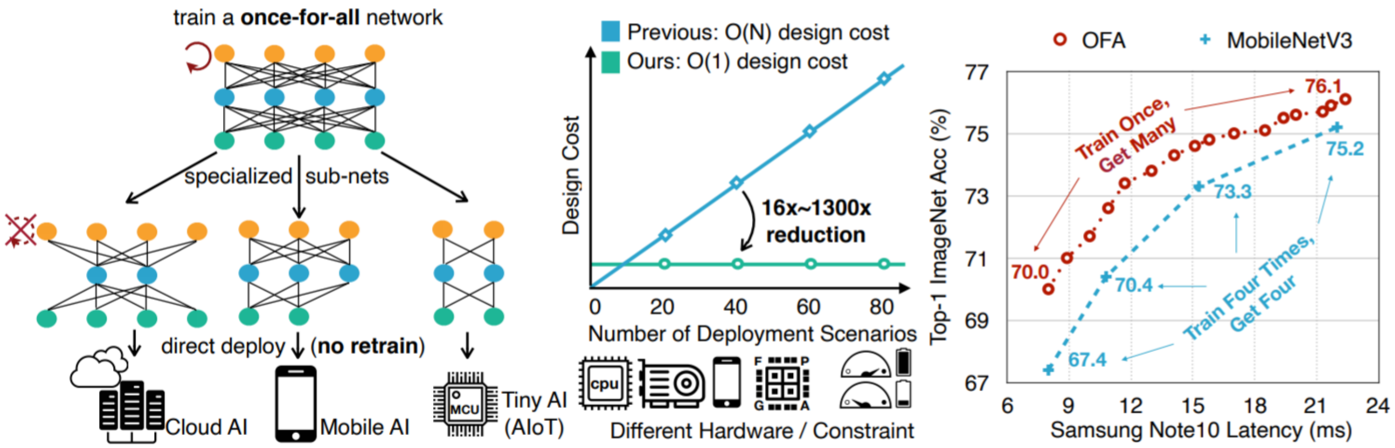
8-bit quantization approximates floating point values using the following formula.

$$real\_value = (int8\_value - zero\_point) \times scale$$

The representation has two main parts:

- Per-axis (aka per-channel) or per-tensor weights represented by int8 two’s complement values in the range [-127, 127] with zero-point equal to 0.
- Per-tensor activations/inputs represented by int8 two’s complement values in the range [-128, 127], with a zero-point in range [-128, 127].

www.tensorflow.org



arxiv.org/abs/1908.09791

#ISSLearningFest

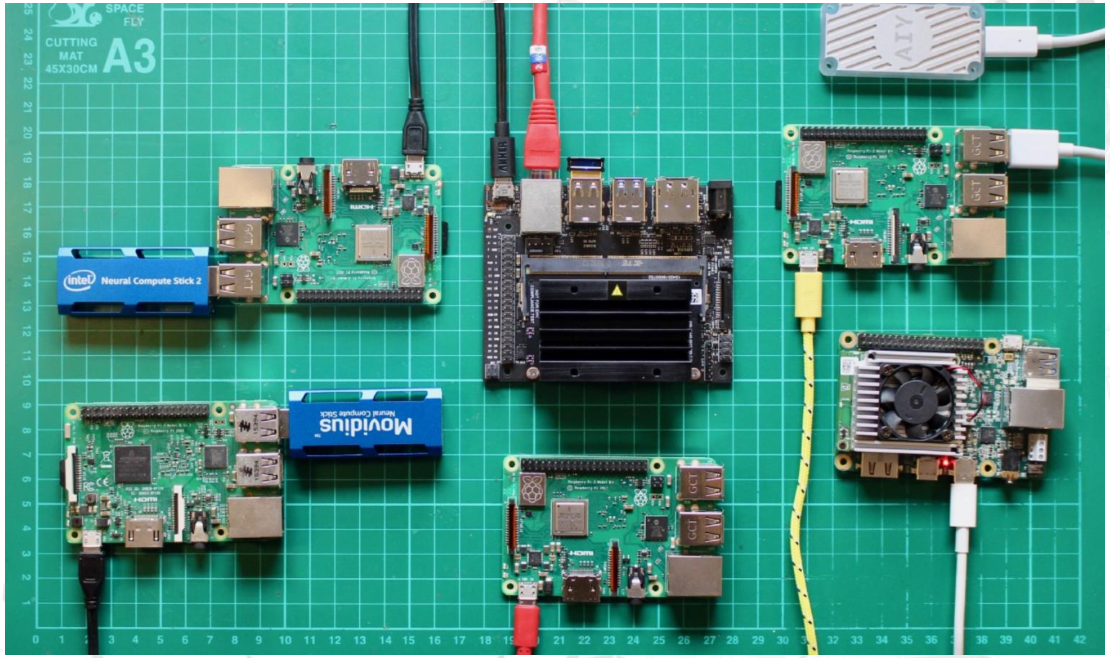
Bridging the Gap (2)

Shrink Deep Neural Networks

Power-up Embedded Things

- Boost using a GPU or TPU, or FPGA add-on
- 4x+ to 10x speedup

Optimise Tools & Workflows



Board	MobileNet v1 (ms)	MobileNet v2 (ms)	Idle Current (mA)	Peak Current (mA)	Price (US$)
Coral Dev Board	15.7	20.9	600	960	$149.00
Coral USB Accelerator	49.3	58.1	470	880	$74.99+$35.00
NVIDIA Jetson Nano (TF)	276.0	309.3	450	1220	$99.00
NVIDIA Jetson Nano (TF-TRT)	61.6	72.3			
Movidius NCS	115.7	204.5	500	860	$79.00+$35.00
Intel NCS2	87.2	118.6	480	910	$79.00+$35.00
MacBook Pro[1]	33.0	71.0	1570	1950	>$3,000
Raspberry Pi	480.3	654.0	410	1050	$35.00

[1] The MacBook Pro takes a +20V supply, all other platforms take a +5V supply.

medium.com/@aallan/benchmarking-edge-computing-ce3f13942245

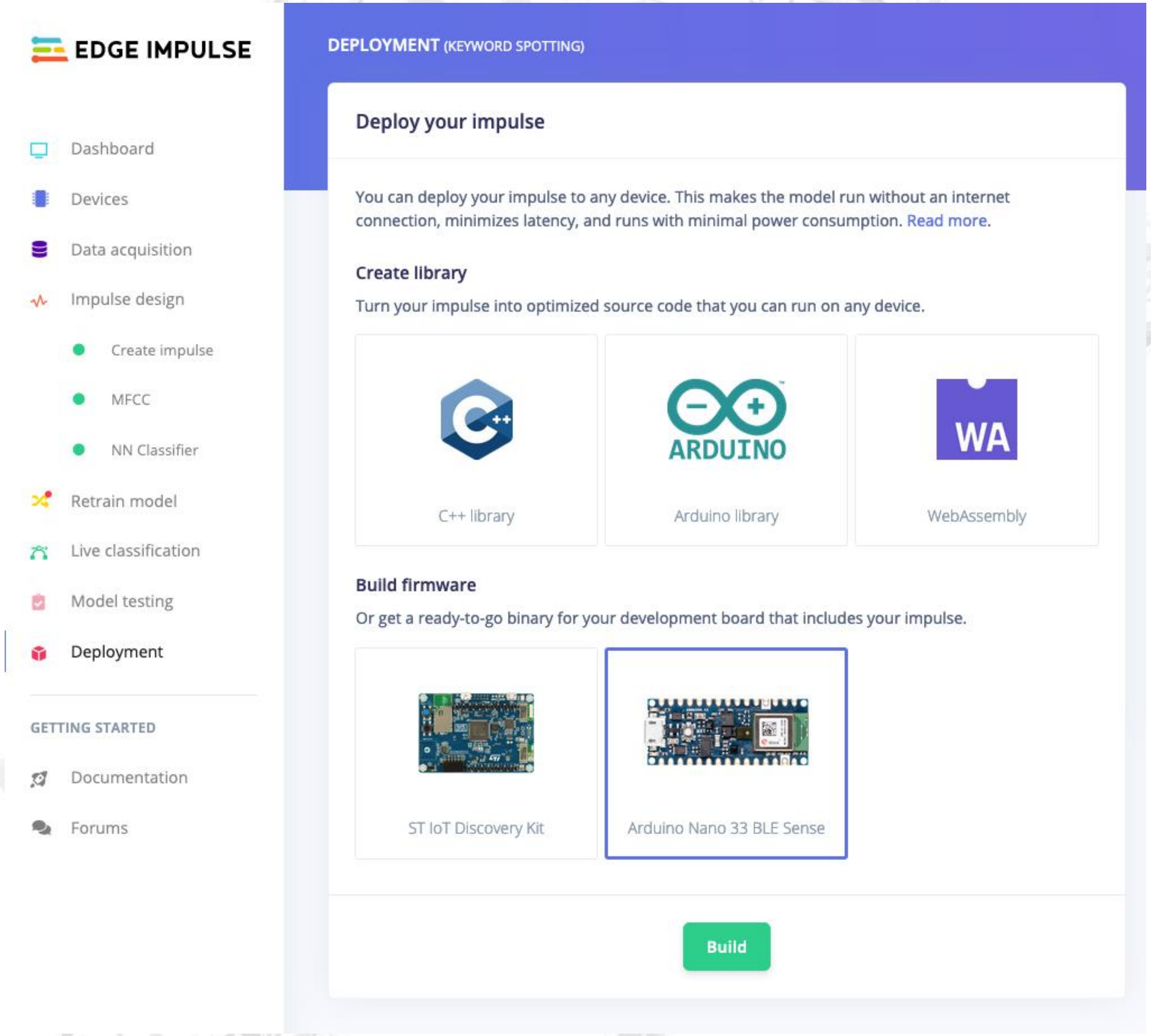
Bridging the Gap (3)

Shrink Deep Neural Networks

Power-up Embedded Things

Optimise Tools & Workflows

- ML Pipelines for Embedded Things
    - Edge Impulse, Qeexo AutoML
- Libraries
    - Eloquent TinyML
    - Eloquent Arduino



www.edgeimpulse.com/blog/edge-impulse-brings-ml-to-arduino

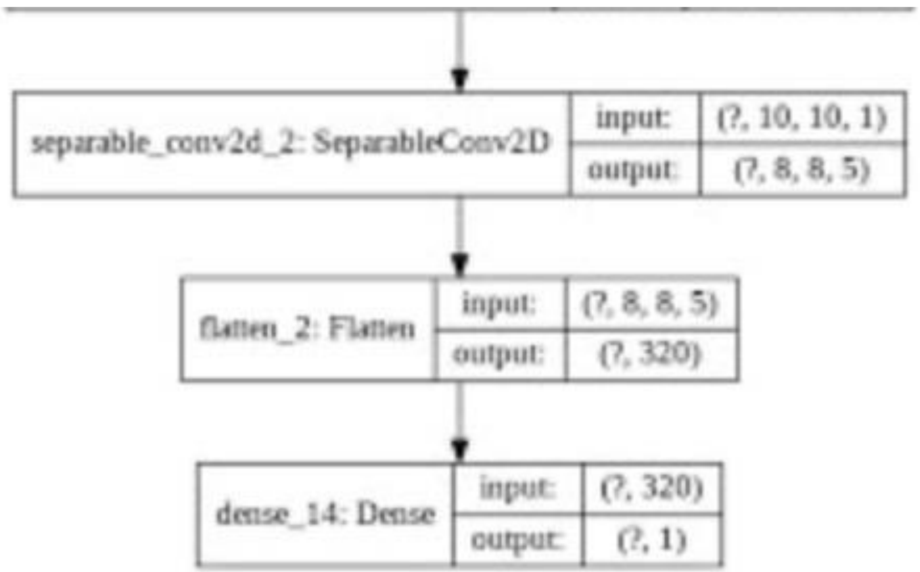
Demo

TinyML on ESP32 Arduino

Mask or Not? TinyML Classifier

1. Train a Convolutional Neural Network on **TensorFlow** to classify if a face image is wearing mask or not
2. Quantize the CNN using **TensorFlow Lite**
3. Deploy to ESP32 using **Arduino IDE**
4. From the ESP32, call the CNN using the **Eloquent Arduino Library**

Code: https://bit.ly/isstinyml



TensorFlow Lite

Demo Design Choices

Image pre-processing done **before** sending to ESP32

- Face detection using OpenCV
- Cropping and resizing to 10x10
- Can offload to another Edge device in a pipelined manner

Bluetooth Serial was used to transmit the 10x10 input to the ESP32

- Can consider WiFi: MQTT+TLS or HTTP+TLS

10x10 input size was empirically determined

- The resampling helped highlight the large blobs associated with a face mask
- Simple CNN was sufficient

Summary

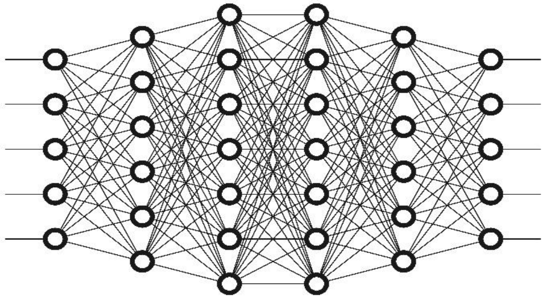
What have we learnt? What are the next steps?

Edge Computing is the New Cloud

More Responsive

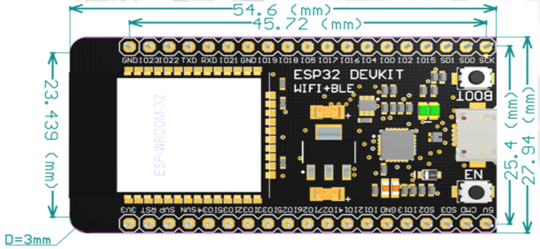
More Resilient

Increased Privacy



www.datanami.com

earlyadopter.com



components101.com

Crossing the Gap

Shrink Deep Neural Networks

Power-up Embedded Things

Optimise Tools & Workflows

Next steps: if you are mulling Embedded ML

Break down the problem into smaller chunks

- Instead of training 1 complex deep neural network to do everything, train multiple simple networks, chain them into a pipeline

Target a mid- or low-range platform (STM32, ESP32, Arduino)

- Raspberry Pi 3 or 4 "high powered", less commercially viable for mass deployment

Evaluate if you must use ML

- Signal processing and filtering can be done using DSPs
- ML is good at finding latent patterns for multi-domain signals, **once** these signals have been properly filtered/ pre-processed

Come take our Grad Cert ☺

Architecting Smart Systems Graduate Certificate

- Build end-to-end software sensing systems such as automated patient care and monitoring, industry 4.0 self-monitoring factories, route-optimising transport systems
- NICF - Architecting IoT Solutions (4 days)
  - IoT protocols, patterns, practices, ML on IoT, Security
- NICF - Designing Intelligent Edge Computing (4 days)
  - Edge computing, orchestration, decision-making, self-learning
- NICF - Humanizing Smart Systems (4 days)
  - Combining voice, gesture, vision, wearables into a holistic intelligent system
- Practice Module (10 man-days)
- Stackable to the **Masters of Technology in Software Engineering**

Thank You!

lisaong@nus.edu.sg

Linkedin.com/in/lisaong