

A close-up photograph of a man with dark hair, a beard, and a small hoop earring in his left ear. He is wearing a dark blue t-shirt and is looking down at a white tablet device he is holding in his hands. The background is blurred, showing what appears to be a modern interior space.

arm

Machine Learning on MCUs

uTensor

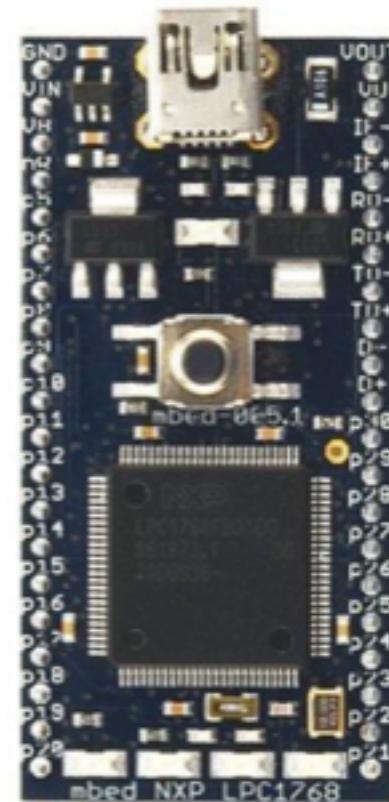
A Mbed Labs Project

A photograph of a large-scale data center or server room. The space is filled with floor-to-ceiling server racks, all illuminated from within by numerous small blue and green lights, creating a glowing, futuristic atmosphere. The perspective is looking down a long aisle between the racks, which leads to a bright, open doorway at the far end of the room. The overall scene conveys a sense of massive computing power and data storage.

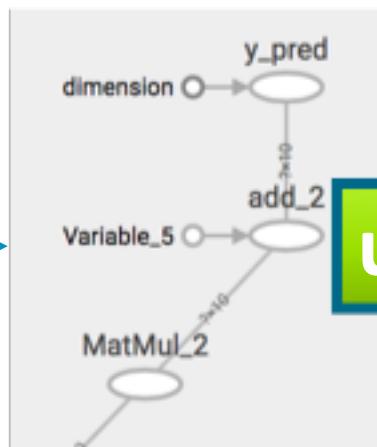
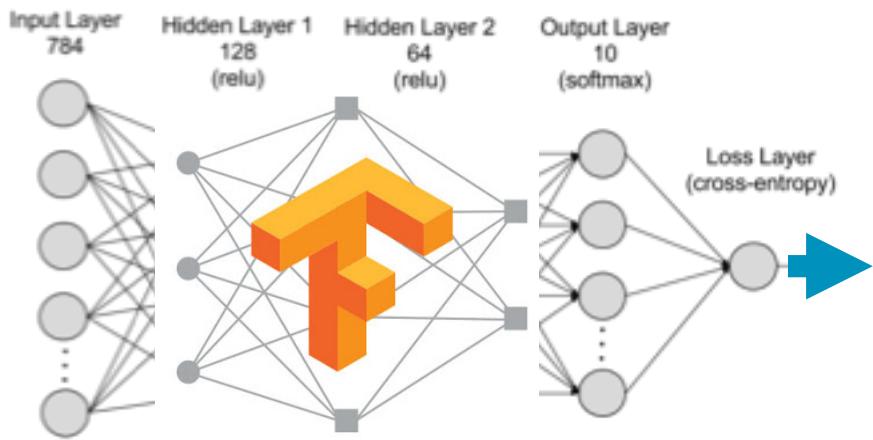
Machine learning

Machine Learning for Microcontrollers

- Runs in <256K RAM
- Runs at ~100 MHz
- TensorFlow Compatible
- Inference Only
- Open source, Apache 2.0 license



uTensor



Graph

uTensor

```
ctx.push(new Requantization_RangeOp(), {"out_mat_p  
...  
//Requantize  
S_TENSOR reqnt_out = ctx.add(new RamTensor<unsigne  
S_TENSOR reqnt_out_min = ctx.add(new RamTensor<flo  
S_TENSOR reqnt_out_max = ctx.add(new RamTensor<flo  
ctx.push(new Requantization_RangeOp<float>("out_ma  
//Requantize  
ctx.add(new RamTensor<float>("out_mat_min_pred", "mat  
ctx.push(new Requantization_RangeOp<float>("out_ma  
//Add  
ctx.add(new RamTensor<float>(), "output_z_pred");  
ctx.push(new AddOp<float>(), {"deqnt_out_pr  
//ArgMax  
ctx.push(new ArgMaxOp<float, int>(), {"output_z_pr
```

C++



Mbed



Tensorflow



5 77984 · 4,3% ALC BY VOL.

100 YEARS

12

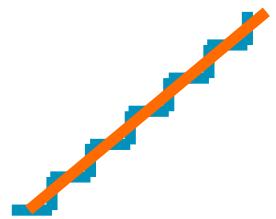
MB1274-D-01
ZT70200300

WST 802.11 b/g/n

16.38

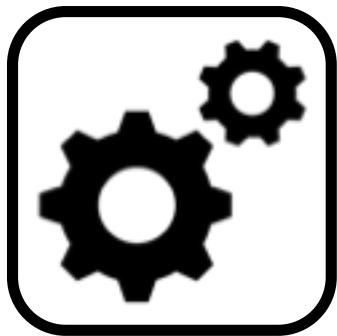
Technologies

Quantization



Float to 8-bit
75% memory saving
Faster Computation

Code Generation



Copy and Paste
Easy Integration

Intermediate Representation



Cross-Framework
Multi-Language

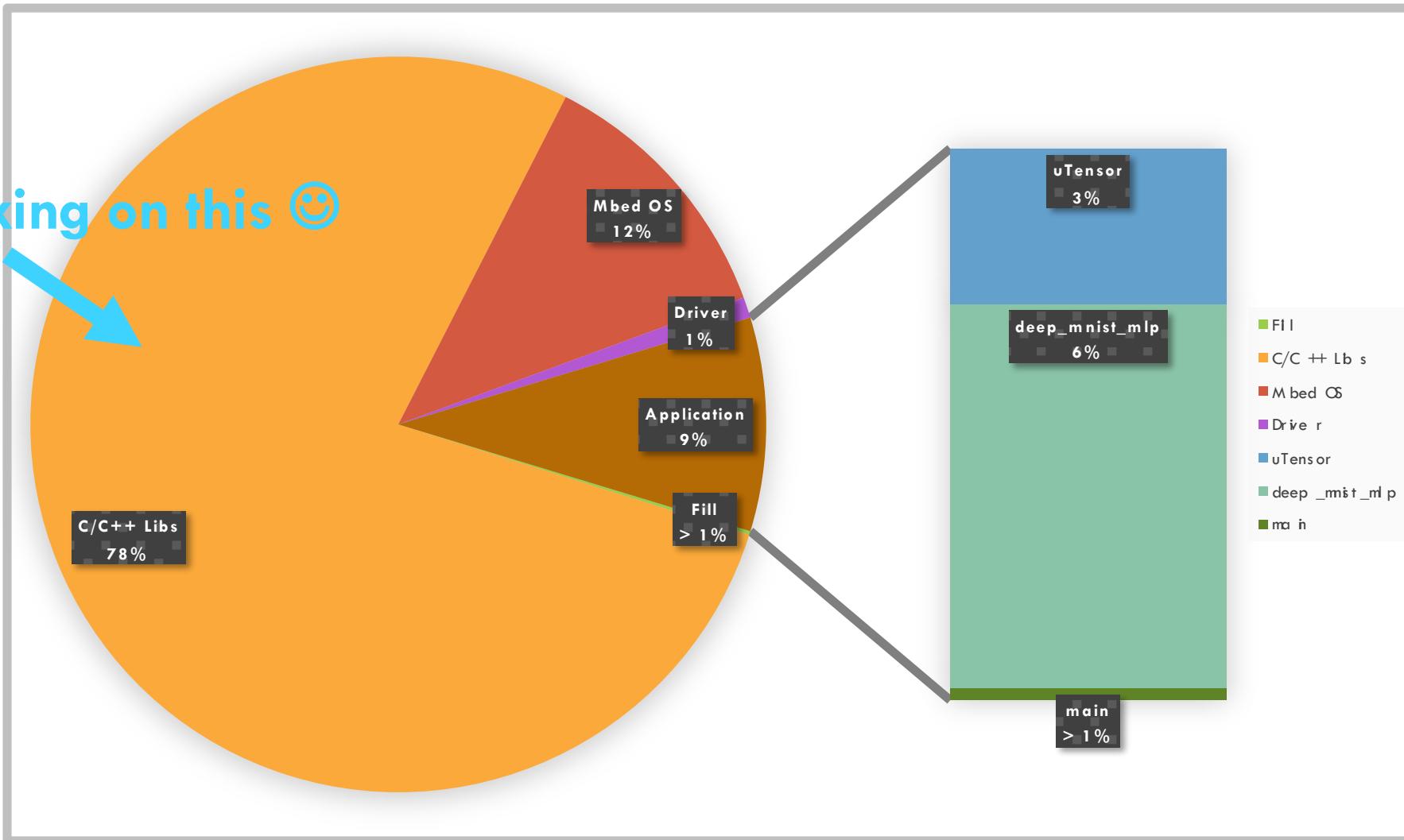
Mbed



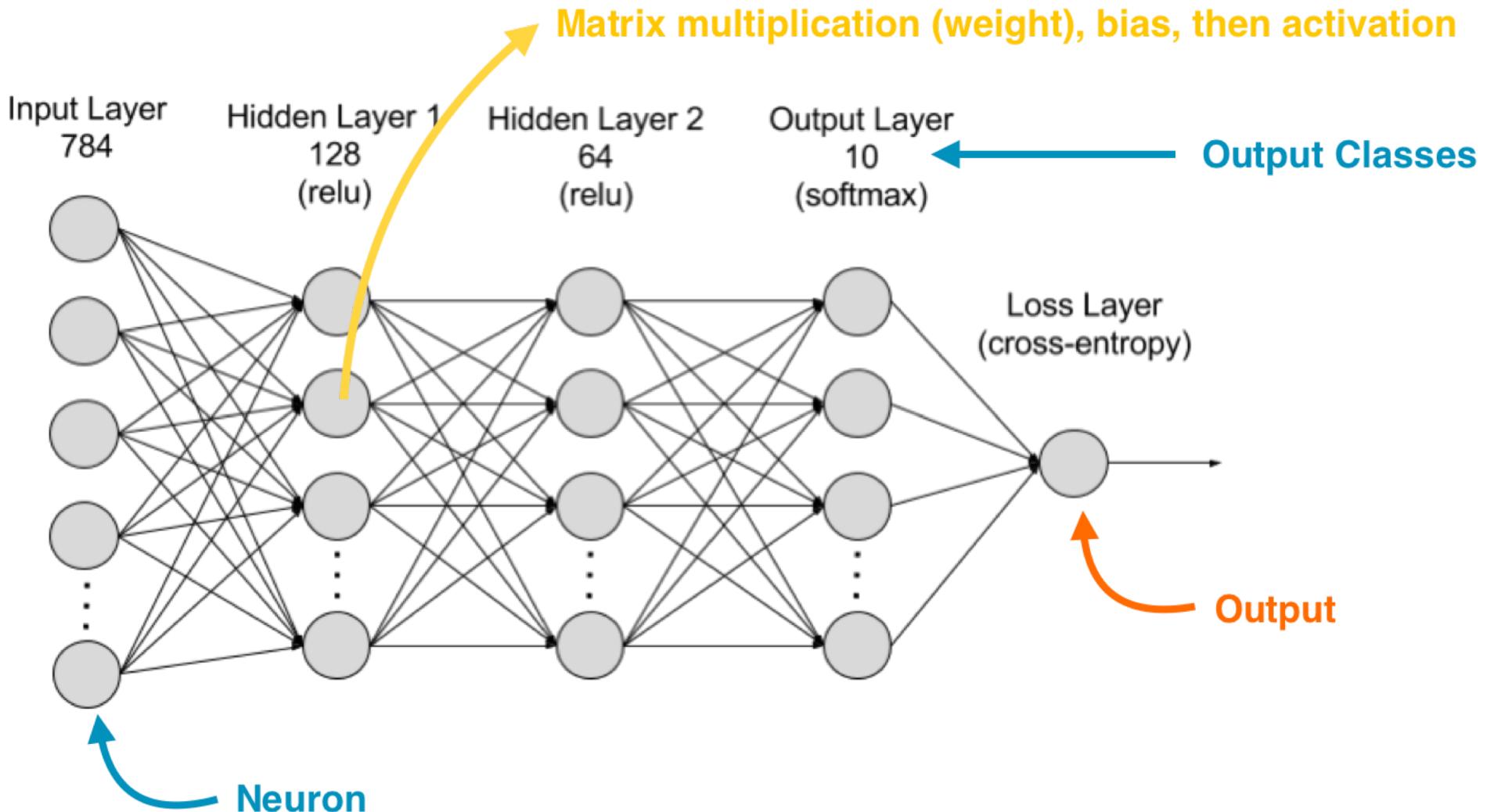
RTOS
CMSIS-NN
Connectivity
Production Ready

Binary

We're working on this 😊

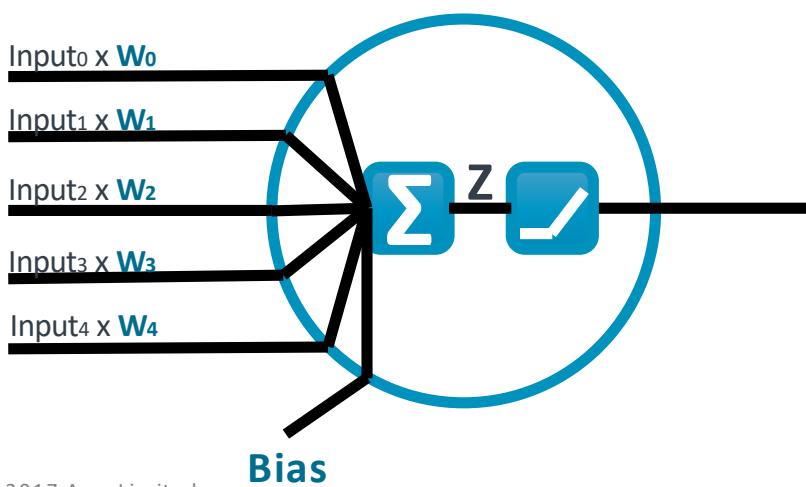


This is a Neural Network



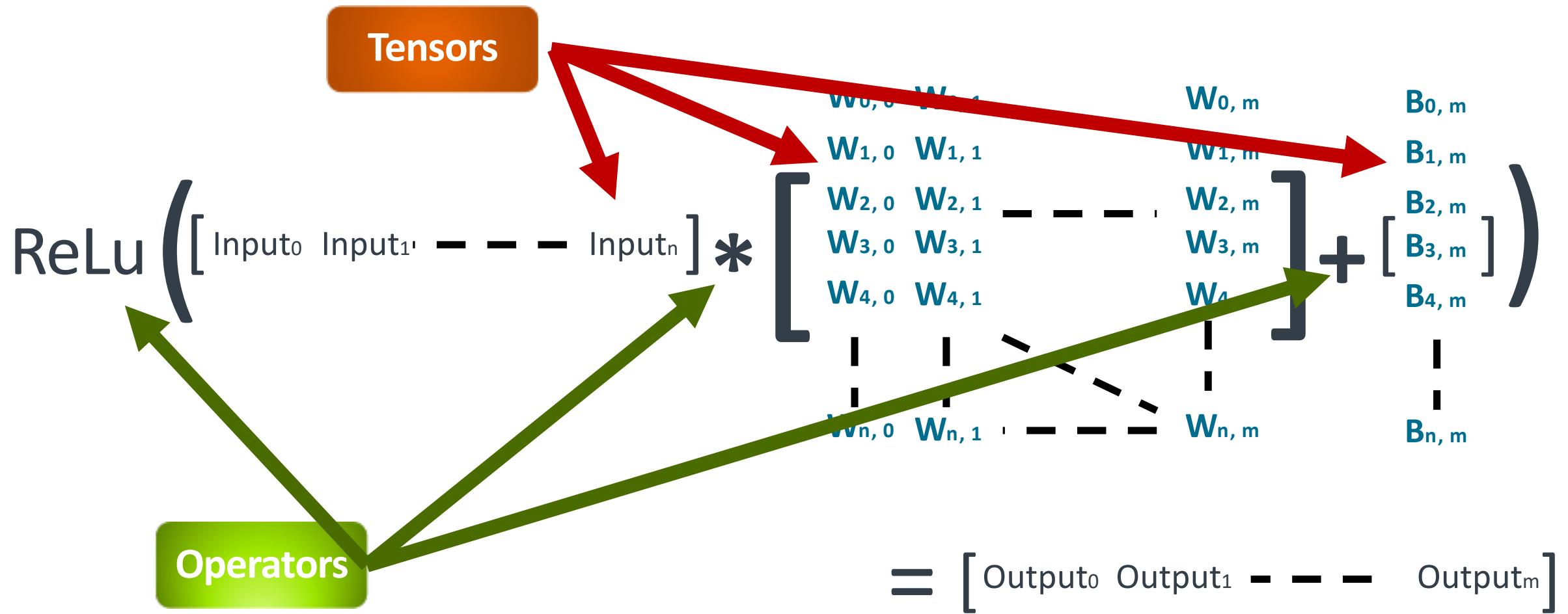
Why Matrix Multiplication

ReLU $([\text{Input}_0 \text{ Input}_1 \dots \text{ Input}_n] * [$



$$\text{Output} = [\text{Output}_0 \text{ Output}_1 \dots \text{ Output}_m] = \text{ReLU} ([\text{Input}_0 \text{ Input}_1 \dots \text{ Input}_n] * [\begin{matrix} \mathbf{W}_{0,0} & \mathbf{W}_{0,1} & & & \mathbf{W}_{0,m} & & \mathbf{B}_{0,m} \\ \mathbf{W}_{1,0} & \mathbf{W}_{1,1} & & & \mathbf{W}_{1,m} & & \mathbf{B}_{1,m} \\ \mathbf{W}_{2,0} & \mathbf{W}_{2,1} & \cdots & \cdots & \mathbf{W}_{2,m} & & \mathbf{B}_{2,m} \\ \mathbf{W}_{3,0} & \mathbf{W}_{3,1} & & & \mathbf{W}_{3,m} & & \mathbf{B}_{3,m} \\ \mathbf{W}_{4,0} & \mathbf{W}_{4,1} & & & \mathbf{W}_{4,m} & & \mathbf{B}_{4,m} \\ \vdots & \vdots & & & \vdots & & \vdots \\ \mathbf{W}_{n,0} & \mathbf{W}_{n,1} & \cdots & \cdots & \mathbf{W}_{n,m} & & \mathbf{B}_{n,m} \end{matrix}] + [\mathbf{B}_{0,m} \mathbf{B}_{1,m} \mathbf{B}_{2,m} \mathbf{B}_{3,m} \mathbf{B}_{4,m} \cdots \mathbf{B}_{n,m}])$$

Tensors and Operators

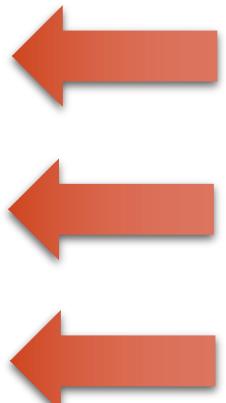


Execution

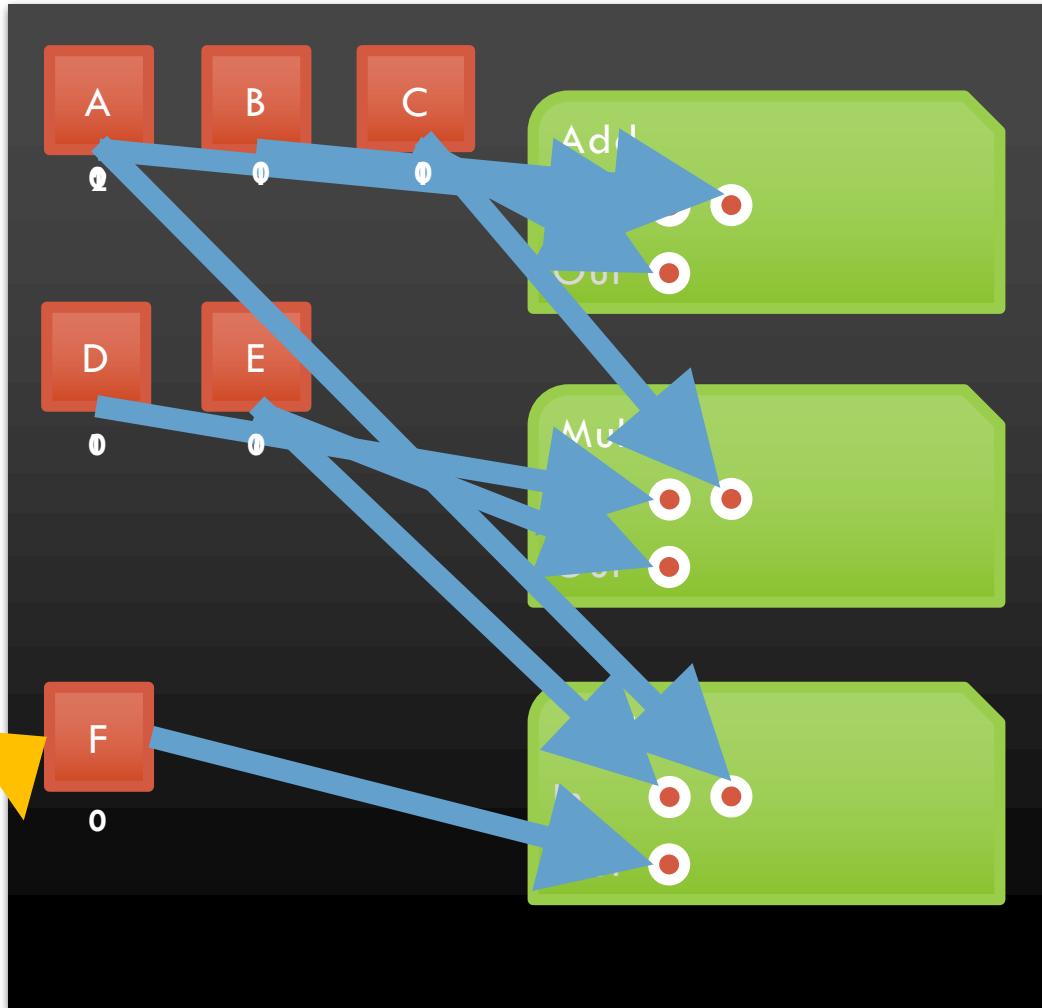
$A + B \Rightarrow C$

$D * C \Rightarrow E$

$A + E \Rightarrow F$



Shared Pointer

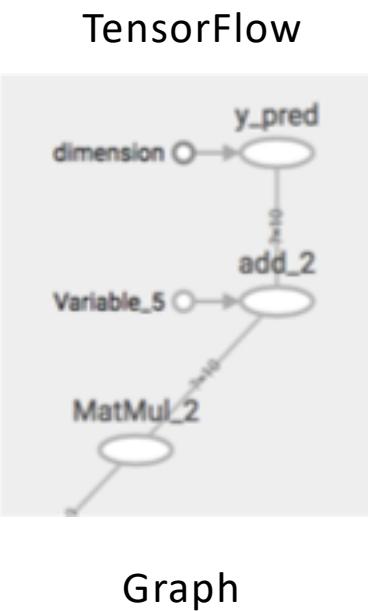


Code generator

A Python Tool

Turns Graph into C++ source

- Graph Traversal
- Optimizer
- Template Engine



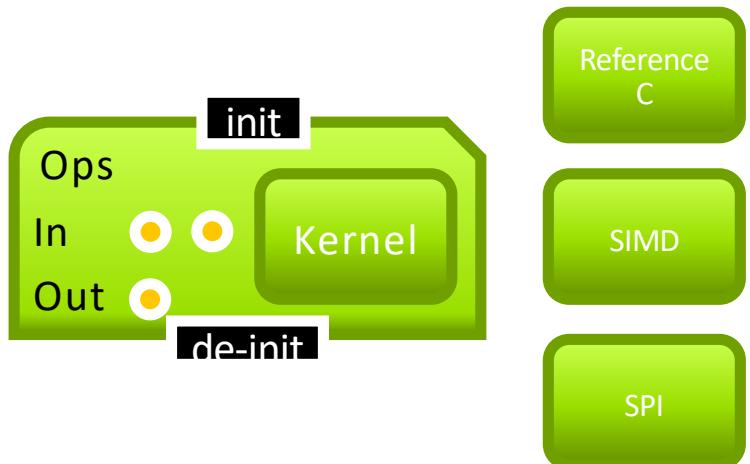
Mbed

```
ctx.push(new Requantization_RangeOp(), {"out_mat_p  
... "req_out_min_pred", "req_out_max_pred"});  
  
//Requantize  
S_TENSOR reqnt_out = ctx.add(new RamTensor<unsigne  
S_TENSOR reqnt_out_min = ctx.add(new RamTensor<flo  
S_TENSOR reqnt_out_max = ctx.add(new RamTensor<flo  
ctx.push(new RequantizeOp(), {"out_mat_pred", "mat  
... "reqnt_out_pred", "reqnt_out_min_pred", "  
  
//dequantize  
ctx.add(new RamTensor<float>(), "deqnt_out_pred");  
ctx.push(new DequantizeOp(), {"reqnt_out_pred", "r  
  
//Add  
ctx.add(new RamTensor<float>(), "output_z_pred");  
ctx.push(new AddOp<float, float>(), {"deqnt_out_pr  
  
//ArgMax  
ctx.push(new ArgMaxOp<float, int>(), {"output_z_pr
```

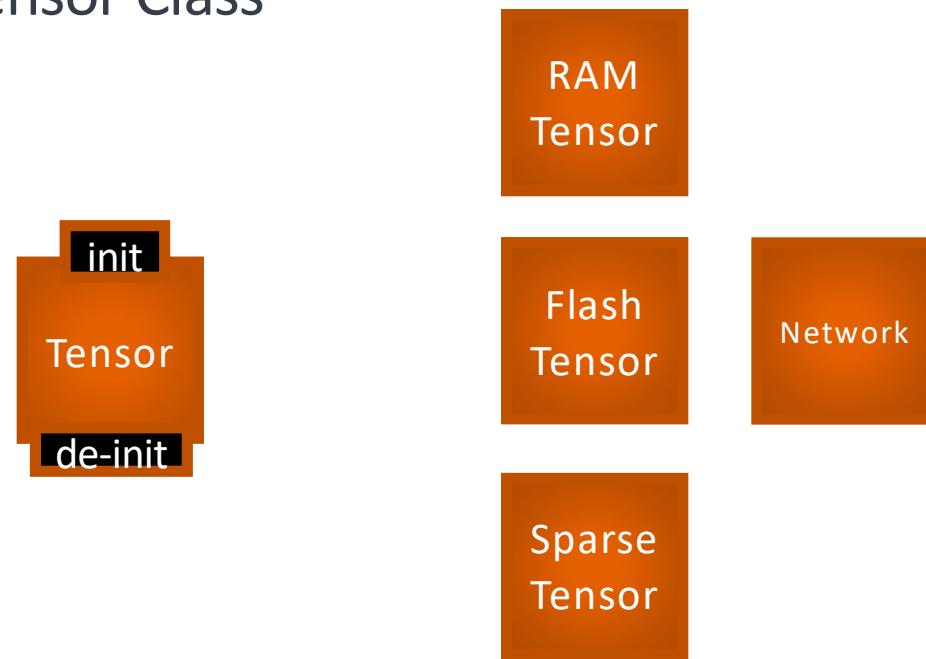
C++

Graph

Operator Class



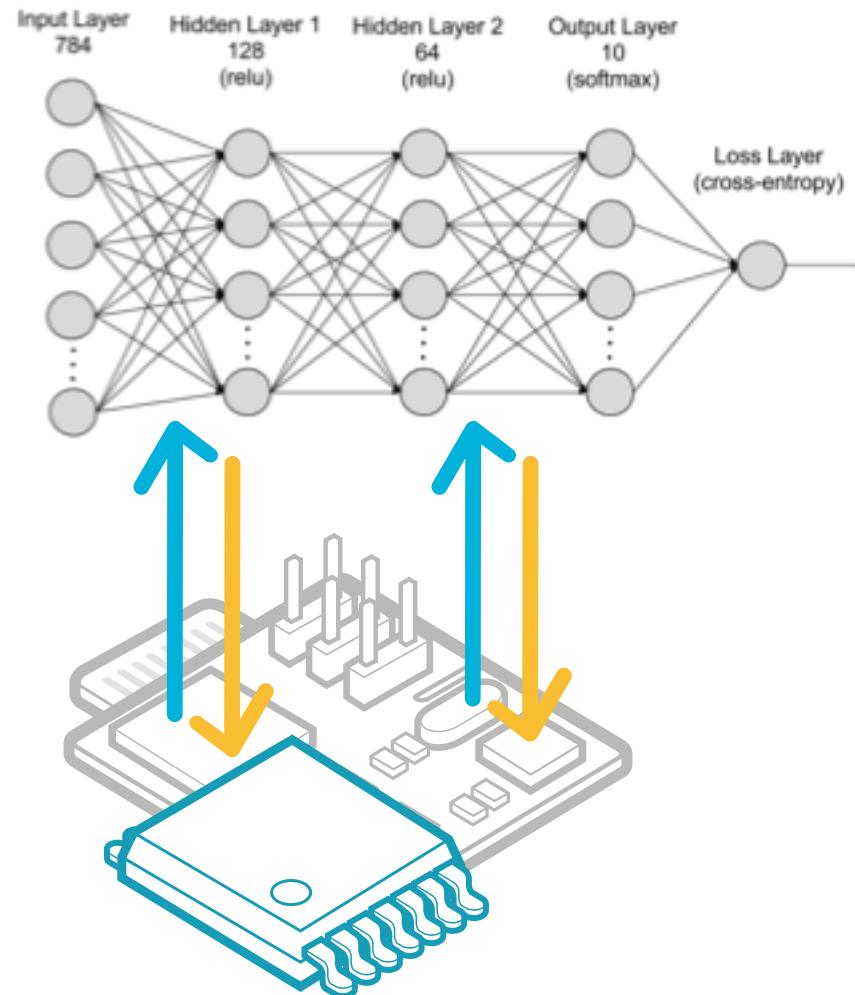
Tensor Class

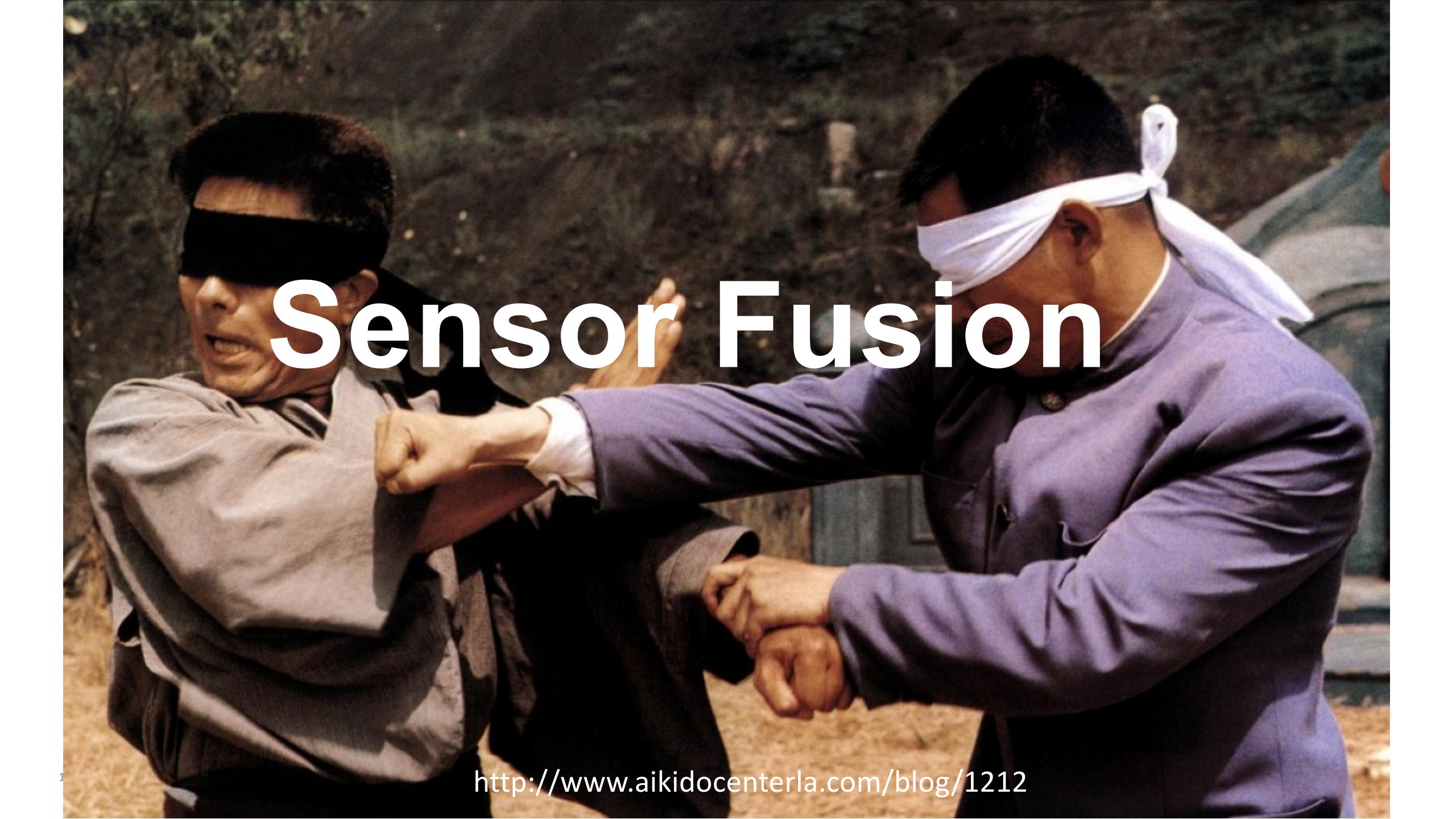


FOTA Graph Update

uTensor's design allows the graph to be embedded in the binary

- Graph is in Firmware
- Firmware Over The Air
- Weights stored in SD or Firmware

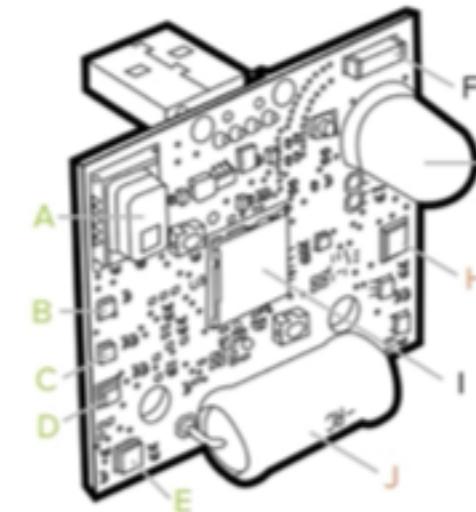
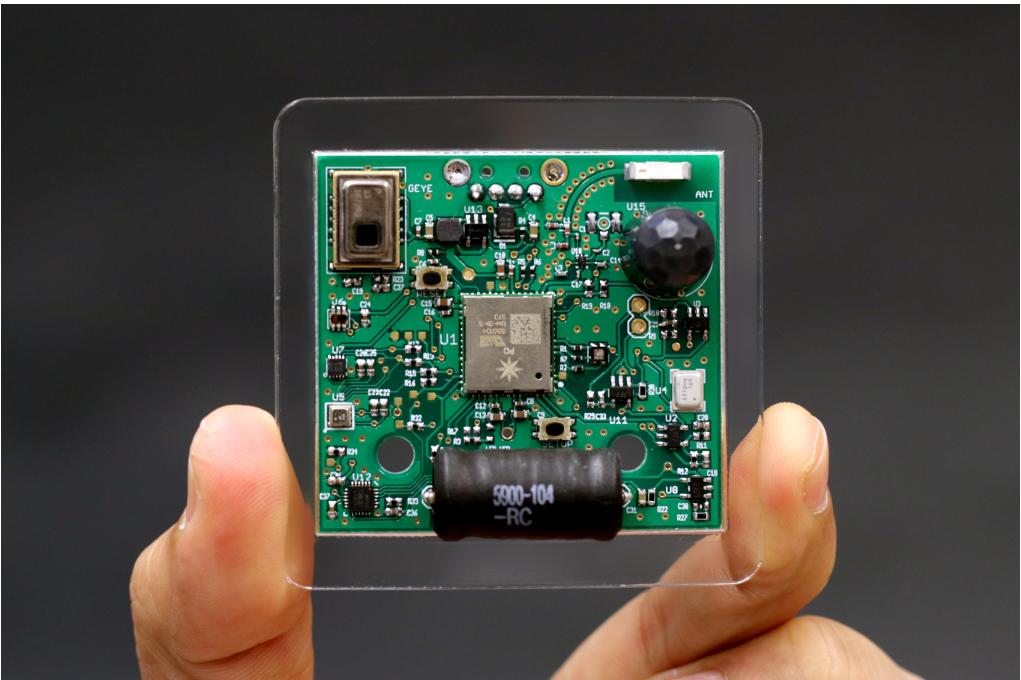


A photograph of two Aikido practitioners in a dynamic pose. One person, wearing a grey gi, is being held by the other person, who is wearing a purple gi. Both individuals have white headbands tied around their eyes, obscuring their vision. They are in a close embrace, with their hands interlocked in a circular motion, illustrating the concept of 'Sensor Fusion' where multiple sensory inputs (in this case, visual and physical touch) are integrated.

Sensor Fusion

<http://www.aikidocenterla.com/blog/1212>

Synthetic Sensors



- A GridEye AMG8833
- B Color / Illumination TCS34725
- C Magnetometer MAG3110F
- D Temp/Baro/Humidity BME280
- E 6-Axis IMU MPU6500
- F 2.4GHz WiFi (RSSI)
- G PIR Motion AMN2111
- H Analog Mic ADMP401
- I Particle Photon STM32F205
- J 100mH Fixed Inductor (EMI)

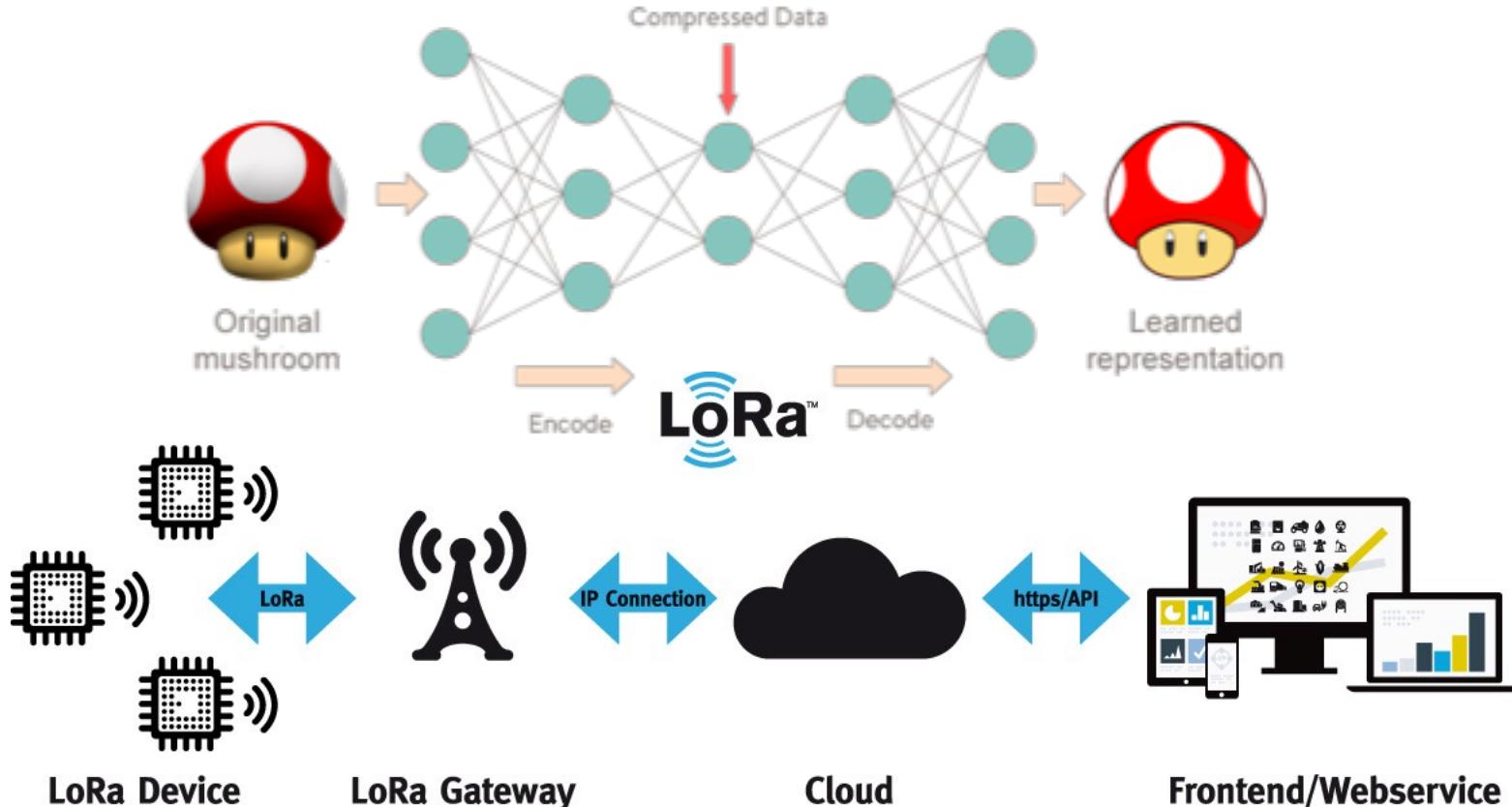
● Digital Components
● Analog Components

<http://www.gierad.com/projects/supersensor/>



Recognized As:
Coffee Grinding

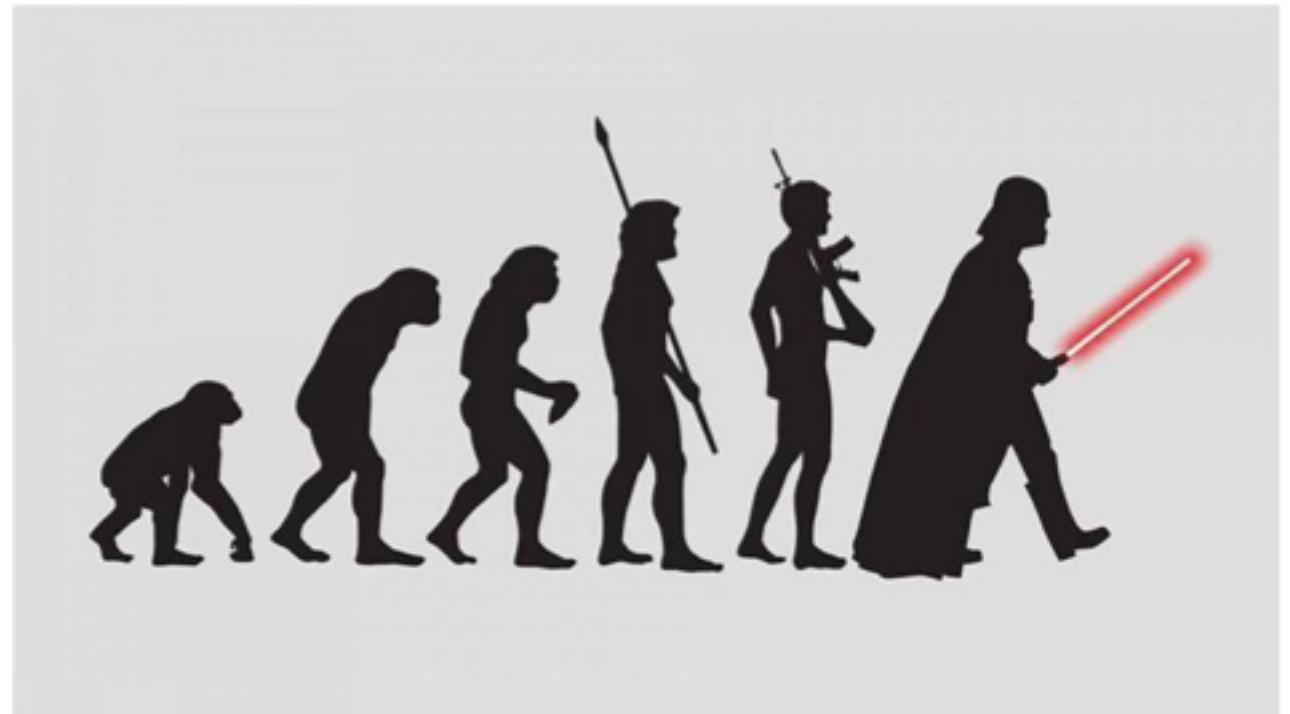
Compression + LPWAN



<http://curiously.com/data-science/2017/02/02/what-to-do-when-data-is-missing-part-2.html>
<https://blog.microtronics.com/lora-and-2g-in-one-module/>

uTensor Timeline

- Test Release
 - Fully Connected, Dropout, Documentation
- Alpha Release
 - Convolution, SD Free
- Beta Release
 - CMSIS-NN



<https://futurism.com/common-misunderstandings-of-evolution-part-2/>



Thank You!