CONTRIBUTORS

KAZAMI HSIEH
ML Researcher, Academia Sinica, Taipei
Startup CTO, Open Source Contributor, Mbed Enthusiast

MICHAEL BARTLING (ARM)
Data Scientist, Mathematician
Mbed Speed Coder

NEIL TAN (ARM)
Physicist, Electrical Engineer
Mbed Mad Scientist

DBOY LIAO
Data Scientist, Piniko, Taipei
Mathematician, Economist, Software Developer
MCU EXAMPLE

- 100 MHz
- 320 KB SRAM
- 1.5 MB Flash
- 112 μA/MHz (1.7 V to 3.6 V)
HELLO WORLD

- MNIST Dataset
- 3-layer MLP
- Less than 256Kb of RAM

https://mxnet.incubator.apache.org/tutorials/python/mnist.html
HOW
DESIGN

• Quantization
  • Float to 8-bit
  • 75% Memory reduction
  • See Pete Warden’s Blog
• Dynamic Tensor Allocation
• Graph to C++ Code Generation
• Optimized Memory Usage
• Mbed for rapid development
WHAT – UTENSOR WORKFLOW

1. Collect Data
2. Train Graph
3. Generate C++
4. Compile Binary
5. Flash
6. Run
CODE GENERATOR

Graph

```
(ut) $ utensor-cli tests/deep_mlp/quant_mnist.pb
...
  saving constants/quant_mnist/OutputLayer_Variable_1_0.idx
  saving constants/quant_mnist/y_pred_dimension_0.idx
  Generate header file: models/quant_mnist.hpp
  Generate source file: models/quant_mnist.cpp
```

Parameters

CPP files
BINARY SIZE

C/C++ Libs 78%
Mbed OS 12%
Driver 1%
Application 9%
Fill > 1%

We’re working on this 😊

C/C++ Libs
Mbed OS
Driver
Application
Fill
main
uTensor
deep_mnist_mlp
We’re working on this 😊
FLEXIBLE CLASSES

Operator Class

Tensor Class

Ops
In
Out
init
de-init

Kernel

Reference C
Remote

SIMD
SPI

init
de-init

Tensor

RAM Tensor
Flash Tensor
Sparse Tensor
Network
WHY
WHY DEVICE AI

• Sensor Fusion
• Lora and NB-IoT
• Long Stand-By Time
• Low Cost
• Novel Use-Cases

Synthetic Sensor: http://www.gierad.com/projects/supersensor/
THANK YOU!

- PR Welcome!
- Embedded Developers
- Data Scientists

[GitHub Repository](https://github.com/uTensor)