Title: Image Recognition on 750 microamps and 100mS Inference Time
Author: Dr. Gopal Raghavan

Types of items to be discussed: Edge Intelligence, tinyML, image recognition,
Relevance to tinyML, Direct implementation of a CNN on an embedded MCU from Eta Compute

Eta Compute demonstrates exceptional innovation by the design and implementation of a complex neural networks on unique ultra-low power sensor node processor based with dual core ARM Cortex-M3 plus DSP SoC and continuous frequency voltage scaling. This is a significant step in the realization of intelligent and power efficient tinyML edge nodes, bringing a much needed improvement in computational efficiency at low power. For instance, a previous ARM publication describes the implementation of the same CNN on a Cortex-M7 running at 216MHz inferencing at 30mJ per image while our implementation consumes only 0.6mJ per image, a 50X reduction of energy consumption.

Neural Networks continue to gain interest for deployments in IoT and other mobile and edge devices. Adding intelligence at the network edge enables a significant reduction in unnecessary data transfer which is much more efficient from a power perspective, and further enables the system to focus in on the data which contains the most important and time sensitive information. Having intelligence at the edge also reduces latency for applications that require a real time response and the ability to adapt to the physical world which is often changing with local conditions.

The demonstration is significant because of:

- Resource constraints of flash and SRAM typical of small embedded processors required careful data management and process scheduling
- Limited processor clock speed, again typical of an embedded processor, limits the horsepower available to run the neural network
- Limited power, which is required to maintain a long lifetime of the devices required in deployed IoT networks

Meeting the accuracy and inferencing of published solutions at a significantly reduced energy per classification

Eta Compute's new ECM3532 makes just this generational improvement in power efficiency while implementing a novel version of an already published CNN (L. Lai, N. Suda and V. Chandra, "CMSIS-NN: Efficient Neural Network Kernels for ARM Cortex-M CPU's") to demonstrate a 50X improvement in energy efficiency while maintaining equivalent accuracy for image inferencing on the CIFAR10 database.