Fine-grained energy profiling of deep ConvNets on the Jetson TX1

Crefeda Faviola Rodrigues, Graham Riley, Mikel Luján
The University of Manchester

Introduction

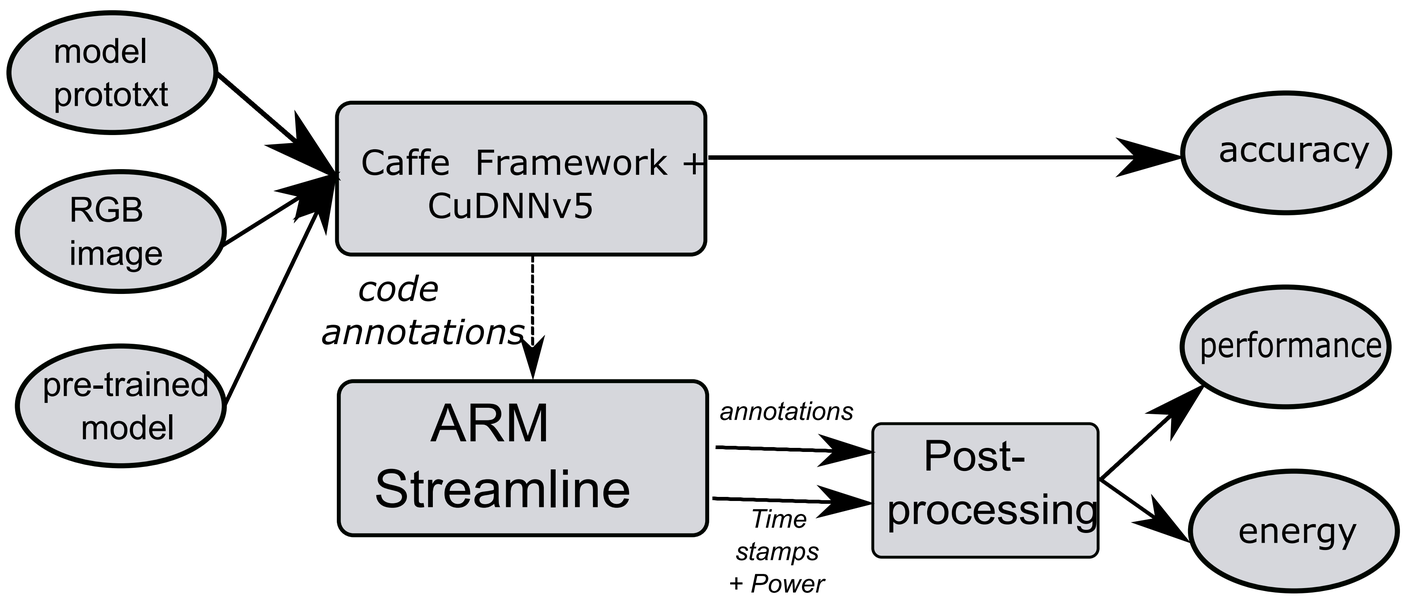
Energy-use is a key concern when migrating current deep learning applications onto low power heterogeneous devices such as a mobile device.

Problem:
Migration is hindered by the lack of tools and methodology to measure power and performance accurately and consistently across devices.

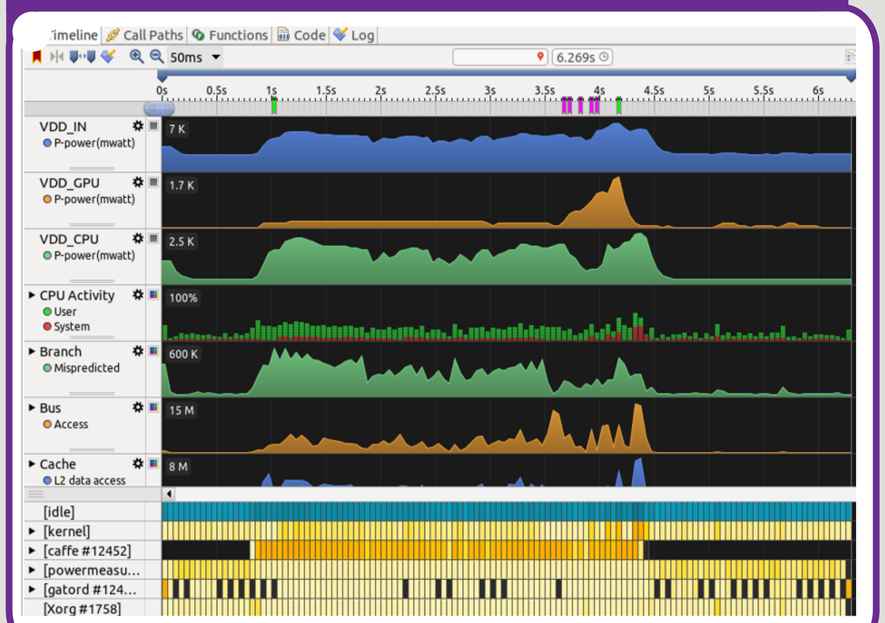
Our solution:
We present a novel evaluation framework for measuring energy and performance for deep convolutional neural networks (ConvNets) using ARM Streamline Performance Analyser integrated with standard deep learning frameworks such as Caffe and CuDNNv5.

Scope:
We apply the framework to study the execution behaviour of SqueezeNet on the Maxwell GPU of the NVidia Jetson TX1, on an image classification task (also known as inference) and demonstrate the ability to measure energy of specific layers of the neural network.

Evaluation Framework



ARM Streamline Annotations



Profiling Results

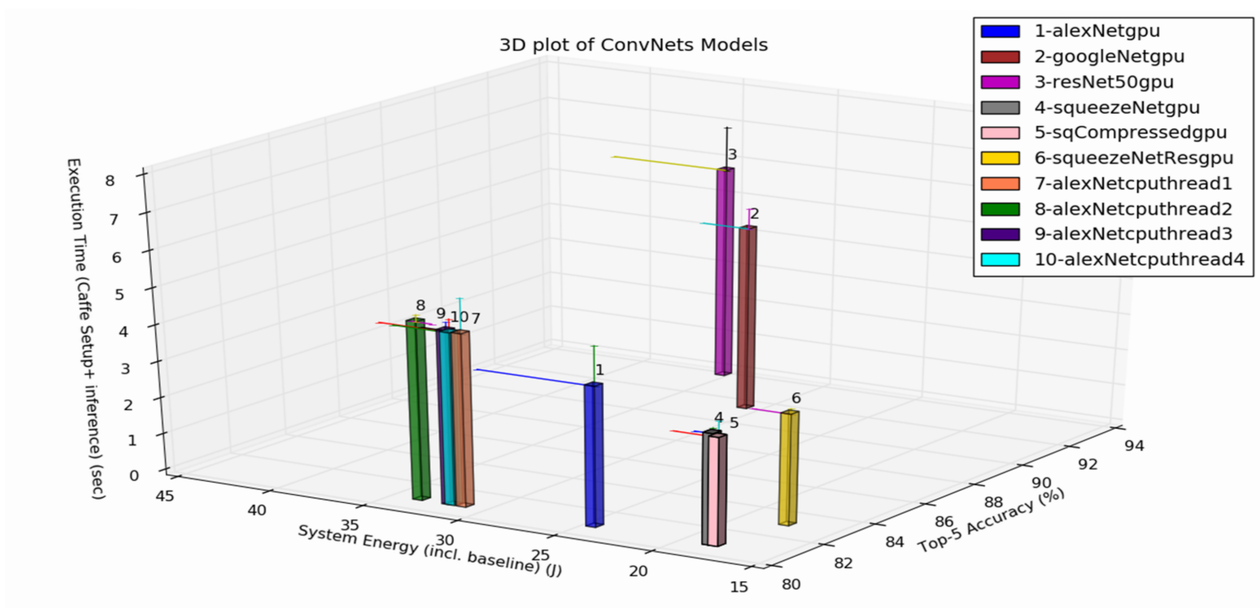
We evaluate several existing ConvNets on the metrics of energy and perfomance.

We report System Energy and Performance for the entire application: Caffe Setup + Inference.

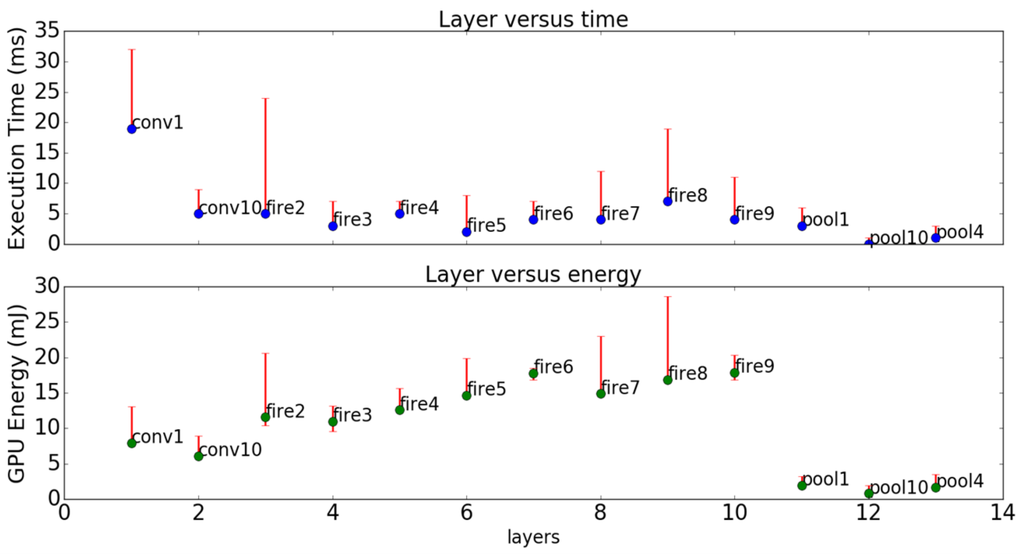
[1] AlexNet evaluated on GPU and CPU respectively.

[2] SqueezeNet + its variants occupy the low energy and low execution time portion of the graph.

[3] Residual Net with 50 layers has a higher accuracy but is also consumes more energy and time.



Per-layer Energy and Performance



Energy consumption and performance (time) of SqueezeNet with inference on the GPU.

Extracted per-layer measurements for conv, pool and fire modules.

✓ Helps understand trends in performance and energy.

✓ Beneficial for deeper analysis work in terms of Bandwidth and FLOPs.

APT Advanced Processor Technologies Research Group

This project is part-funded by ARM and IS-ENES2 project.