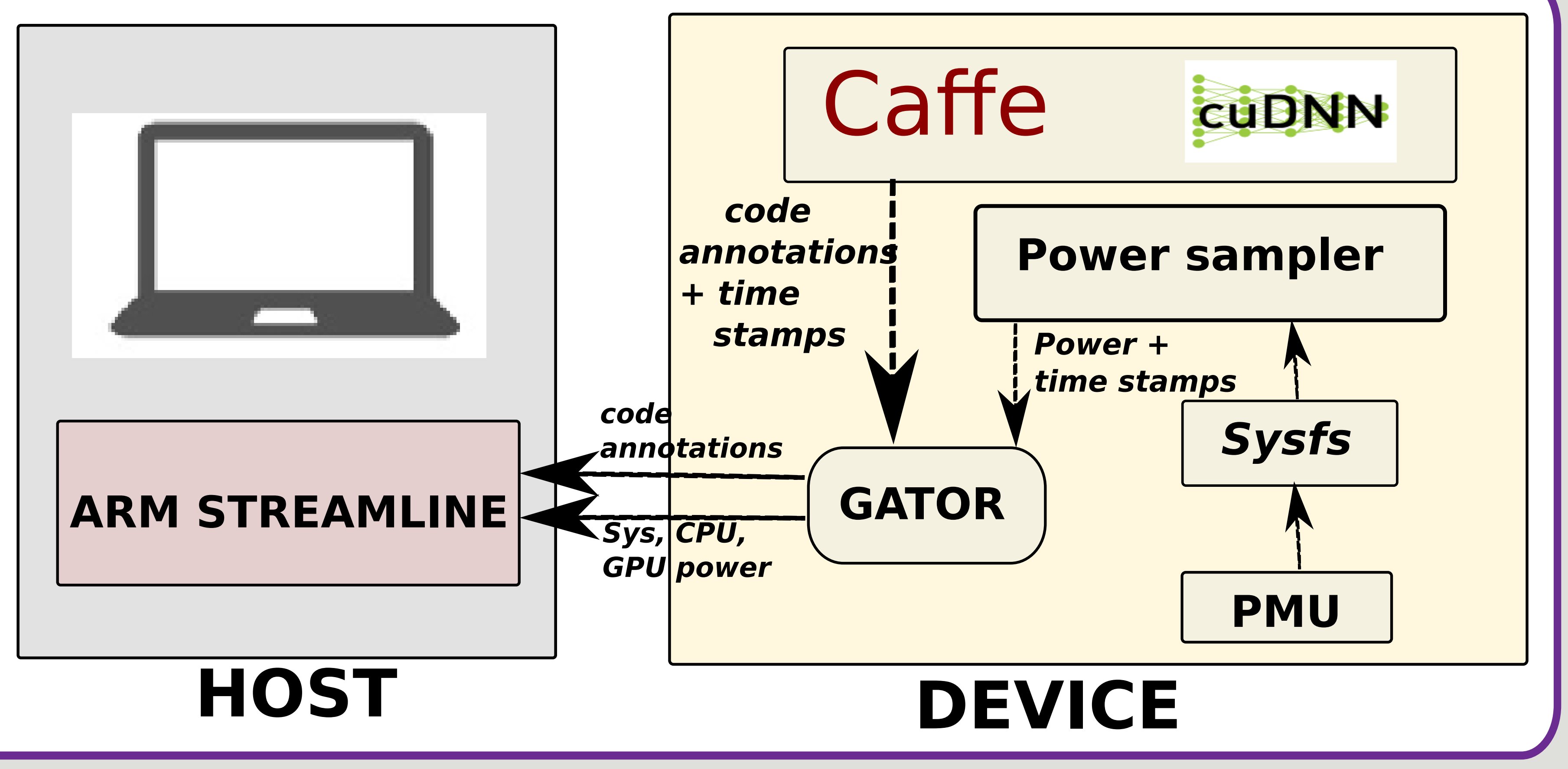
Energy and performance evaluation of ConvNets on low-powered heterogenous systems

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

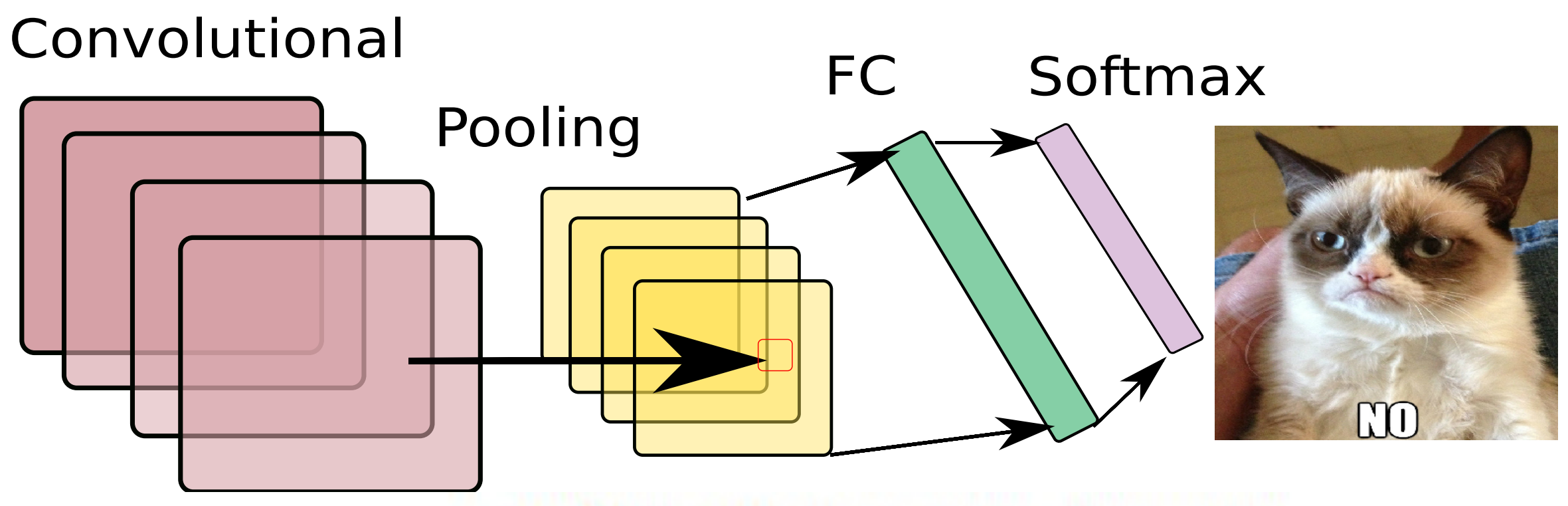
Abstract

Limited power budgets on edge-devices such as mobile systems has resulted in a niche area of research focussed on developing energy-efficient solutions to deep learning problems. Current research shows that there exists a gap to deliver quality power measurements due to the absence of a systematic methodology to enable consistent and accurate power measurements as well as issues with current reporting standards. We present a systematic methodology for measuring energy and performance of Deep Convolutional Neural Networks on low power heterogeneous systems using ARM's Streamline Performance Analyser integrated with standard deep learning frameworks such as Caffe and CuDNNv5. We applied 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 and usefulness to measure energy of specific layers of the neural network.

Evaluation Framework



Convolutional Neural Networks



ConvNet models in the literature

ConvNet	Naming Convention in graphs	Top-5 accuracy (%)	# Layers	Model Size
AlexNet	alexNet	80.3	5	244MB
GoogleNet	googleNet	90.85	22	54MB
Residual Net	resNet50	93.29	50	103MB
SqueezeNet	squeezeNet	80.3	14	5MB
SqueezeNet with Deep Compression	sqCompressed	80.3	14	675.8KB
SqueezeNet with Residual Connections	squeezeNetRes	82.5	14	6.3MB

Profiling Results

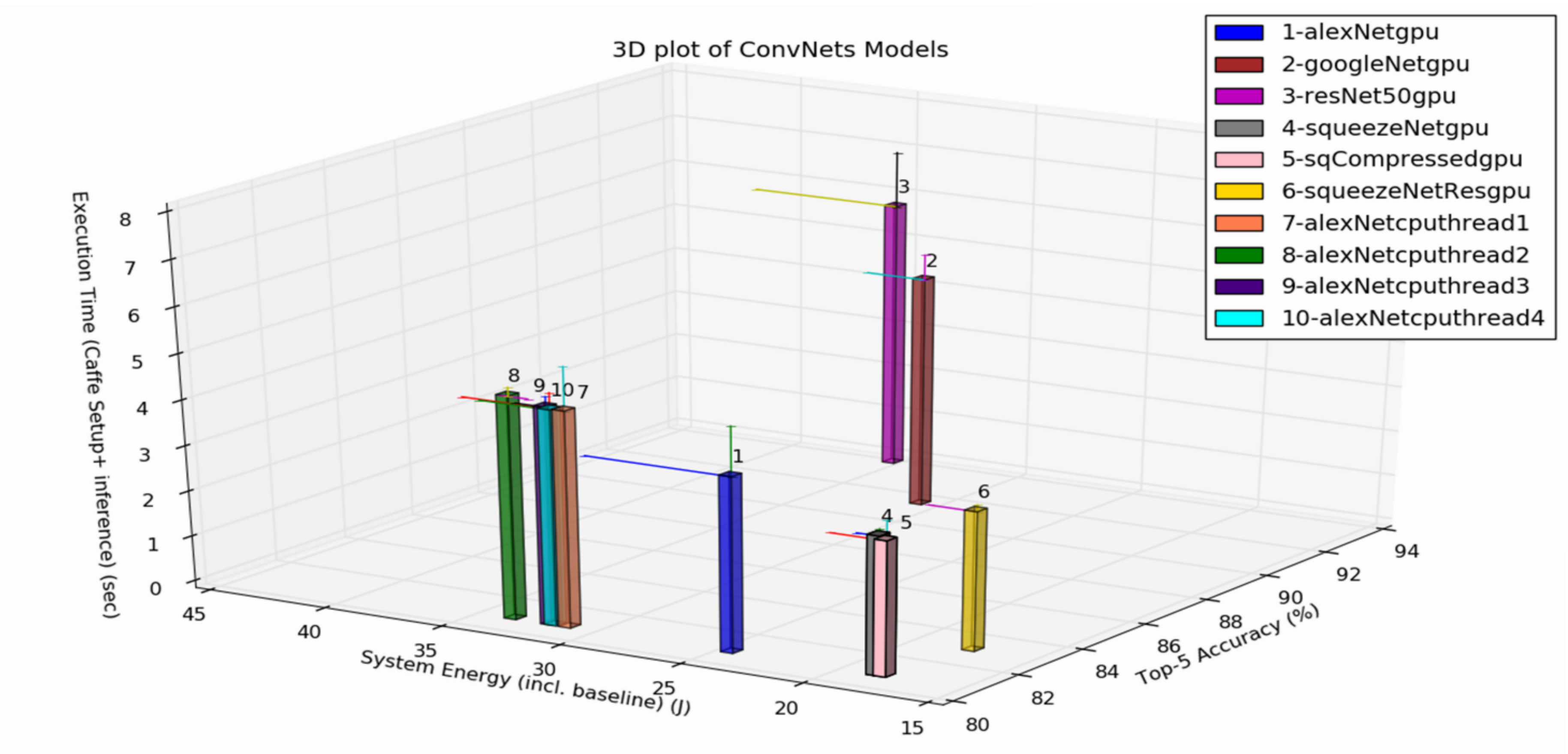
We evaluate several existing ConvNets on the metrics of energy and perfomance.

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

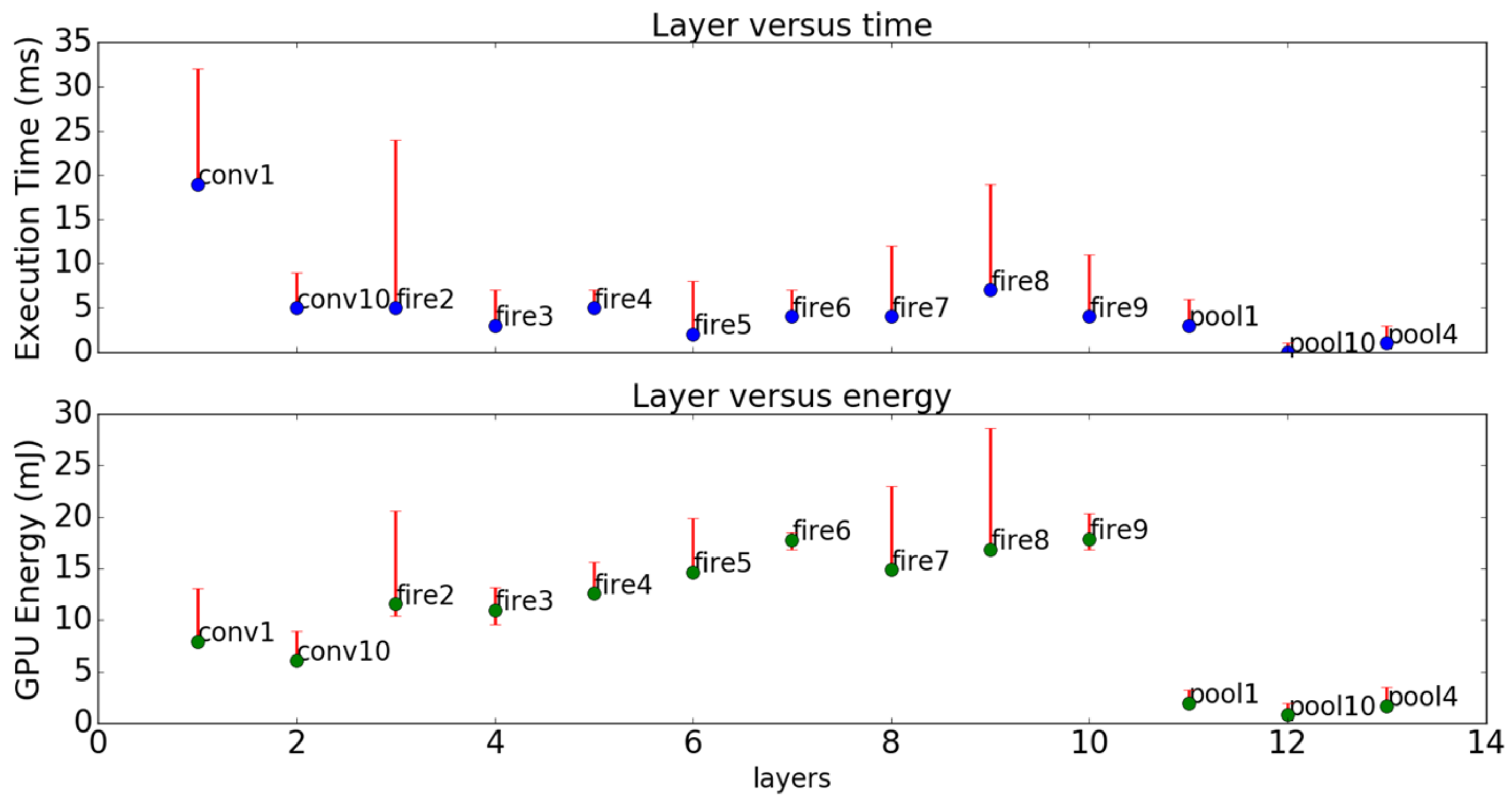
[1] AlexNet evaluated on GPU and CPU respectively.

[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.

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

APT Advanced Processor Technologies Research Group

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING