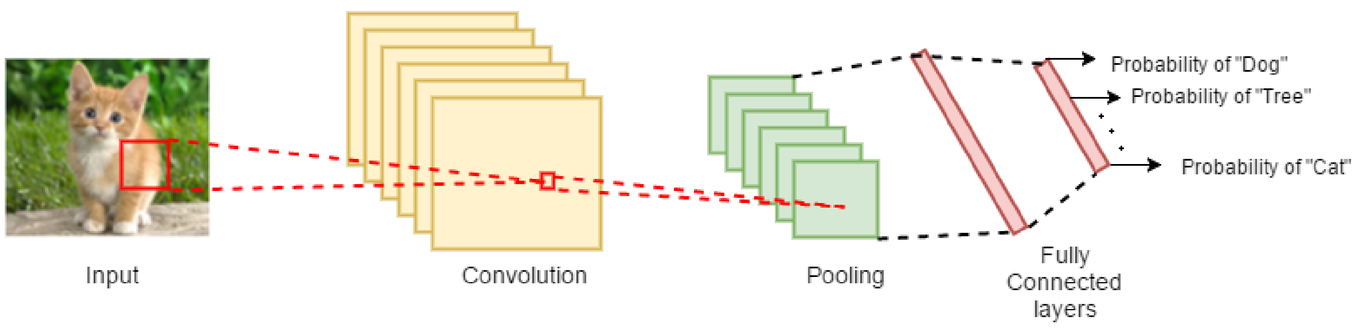
Efficient Execution of Deep Convolutional Neural Networks on Low Power Heterogeneous Systems

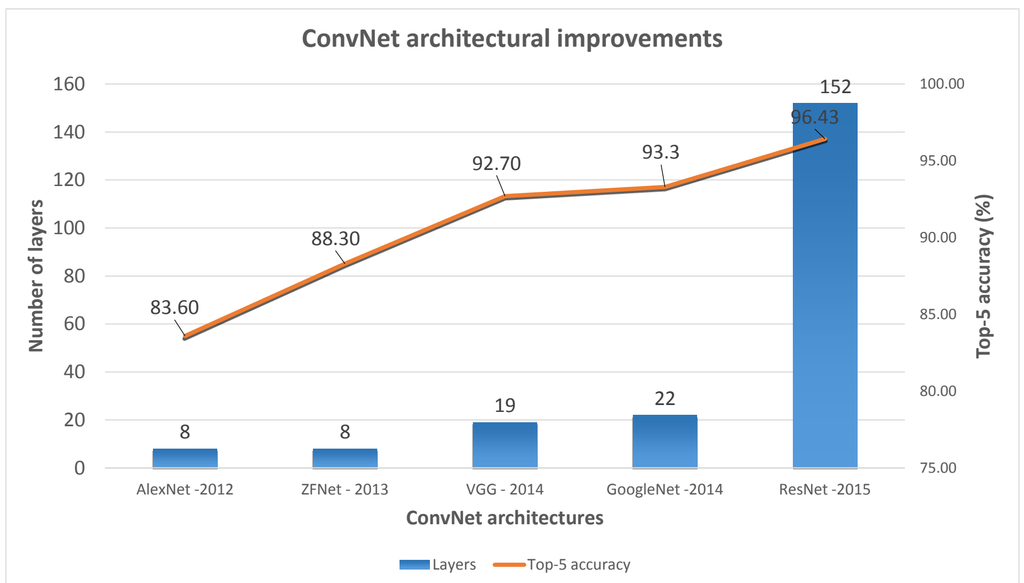
Crefeda Faviola Rodrigues, Graham Riley, Mikel Luján
The University of Manchester, UK

I. Introduction

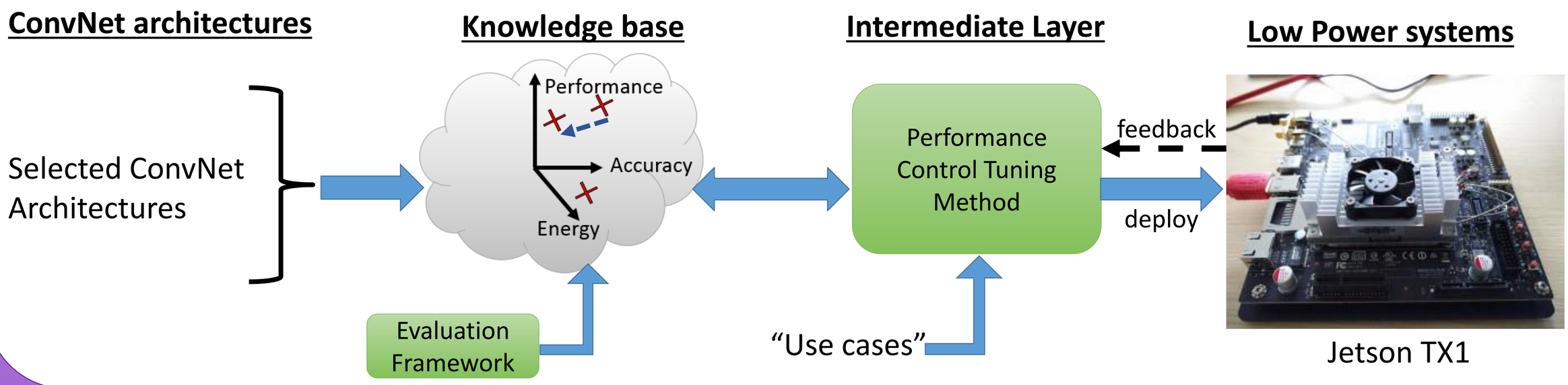
- Deep Convolutional Neural Networks (ConvNets) are algorithms capable of solving a number of computer vision tasks. For example, recognition, localization, detection and segmentation.
- Mobile and wearable devices are now equipped with high-resolution sensors which offer a unique opportunity to build applications that employ these algorithms. These devices impose constraints in terms of computing and memory resources, and mapping these algorithms to best use all possible resources at every clock cycle is still an open research area.
- This project will focus on the following objectives:
  - Understanding the accuracy, performance and energy characteristics of deep ConvNets in the context of current low power heterogeneous systems.
  - Developing methods to exploit this information to enable efficient execution of existing and new ConvNet architectures.

II. Deep Convolutional Neural Networks

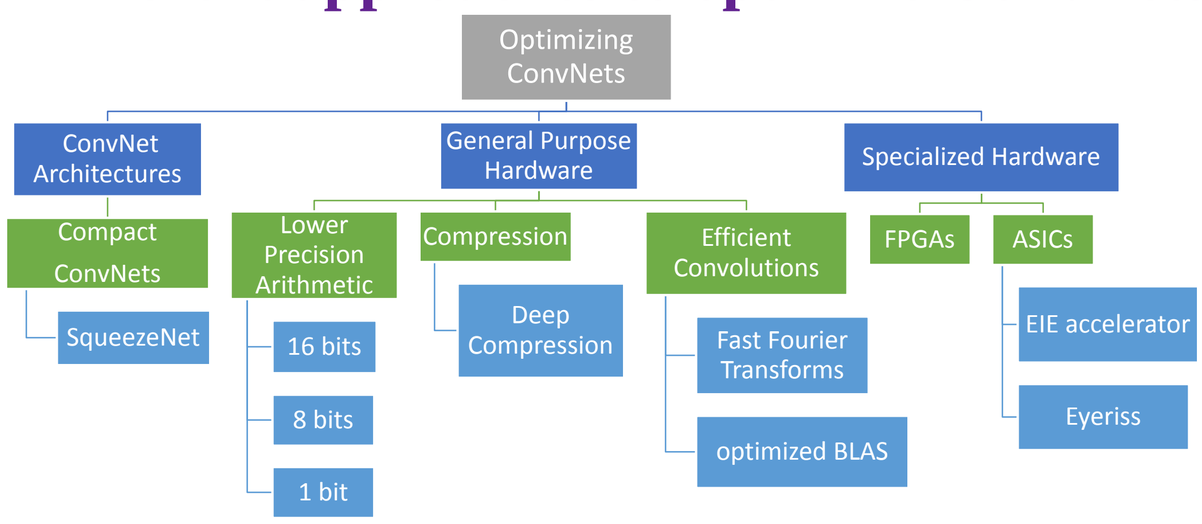




III. Approach



V. Different approaches to optimize ConvNets



VI. Key Research Questions

- What is the ***dynamic trade-off*** of performance, accuracy and energy ?
- Can we develop/train ***improved*** ConvNet architectures?
- What is the best combination of ***optimizations*** for ConvNets in the literature and new ones?

The University of Manchester

APT Advanced Processor Technologies Research Group

ARM

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