Imperial College London

Deploying Deep Neural Networks in the Embedded Space

Dr. Christos-Savvas Bouganis

2nd Workshop on Reconfigurable Computing for Machine Learning (***RCML***)

30 August 2018

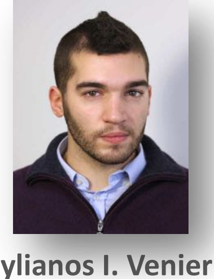
intelligent Digital Systems Lab
Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

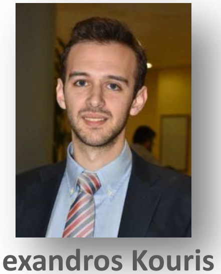
Imperial College London

Intelligent Digital Systems Lab

The team



Stylianos I. Venieris
Machine Learning



Alexandros Kouris
Machine Learning,
Robotics

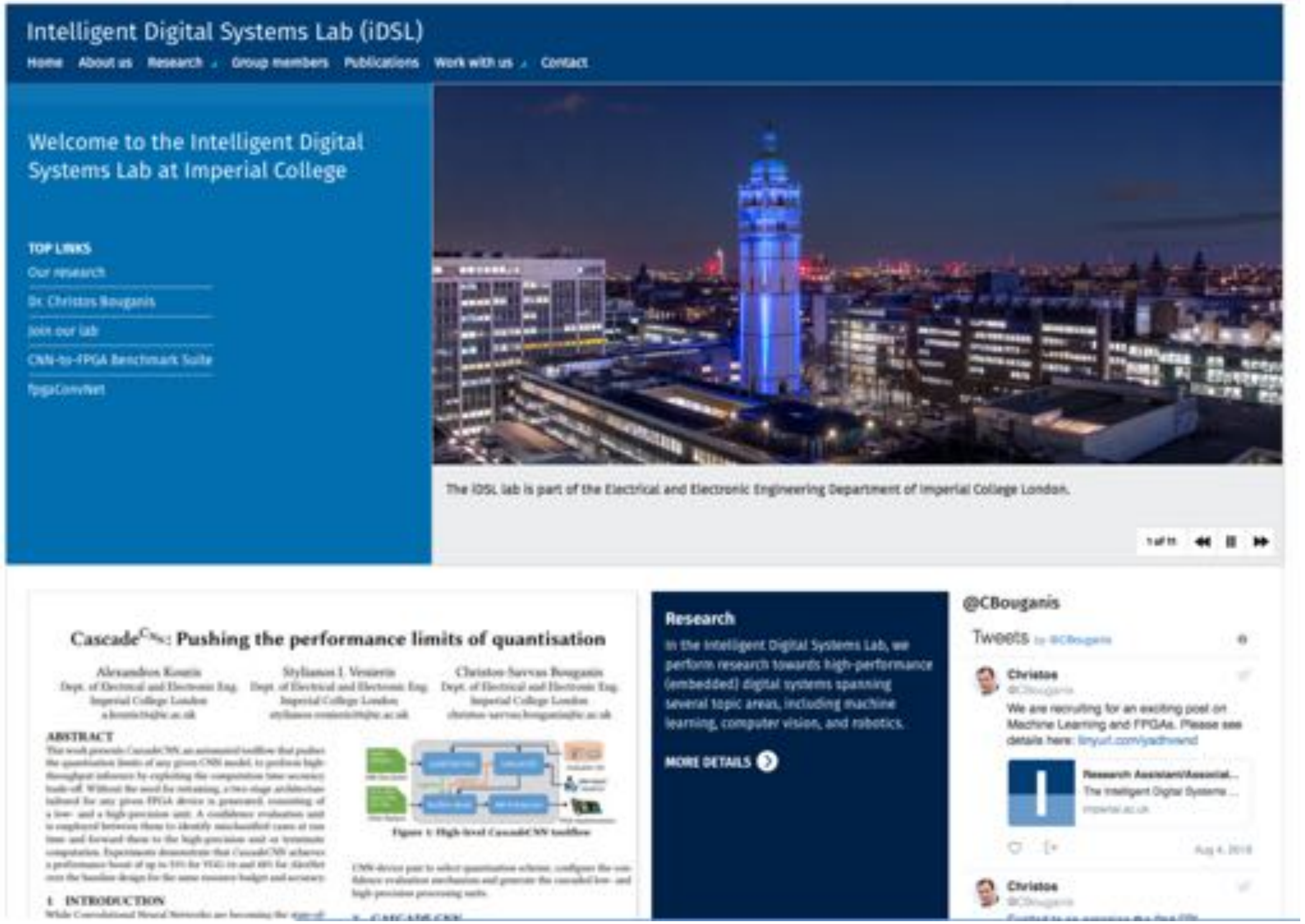
Konstantinos Boikos
Computer Vision,
SLAM

Manolis Vasileiadis
Computer Vision

Mudhar Bin Rabieah
Machine Learning

Nur Ahmadi
Brain-Machine Interface

Christos-Savvas Bouganis
iDSL Lab Director
Imperial College London



Intelligent Digital Systems Lab (iDSL)

Welcome to the Intelligent Digital Systems Lab at Imperial College

TOP LINKS
Our research
Dr. Christos Bouganis
Join our lab
CNN-to-FPGA Benchmark Suite
fpgaConvNet

The iDSL lab is part of the Electrical and Electronic Engineering Department of Imperial College London.

Cascade^{CNN}: Pushing the performance limits of quantisation

Alexandros Kouris, Dept. of Electrical and Electronic Eng., Imperial College London, a.kouris16@ic.ac.uk
Stylianos I. Venieris, Dept. of Electrical and Electronic Eng., Imperial College London, stylianos.venieris10@ic.ac.uk
Christos-Savvas Bouganis, Dept. of Electrical and Electronic Eng., Imperial College London, christos-savvas.bouganis@ic.ac.uk

ABSTRACT
This work presents CascadeCNN, an automated toolflow that pushes the quantisation limits of any given CNN model, to perform high-throughput inference by exploiting the computation time-accuracy trade-off. Without the need for retraining, a two-stage architecture tailored for any given FPGA device is generated, consisting of a low- and a high-precision unit. A confidence evaluation unit is employed between them to identify misclassified cases at run time and forward them to the high-precision unit or terminate computation. Experiments demonstrate that CascadeCNN achieves a performance boost of up to 55% for VGG-16 and 48% for AlexNet over the baseline design for the same resource budget and accuracy.

1 INTRODUCTION
While Convolutional Neural Networks are becoming the state-of-

Figure 1: High-level CascadeCNN toolflow

CNN-device pair to select quantisation scheme, configure the confidence evaluation mechanism and generate the cascaded low- and high-precision processing units.

2 CASCADECNN

Research
In the Intelligent Digital Systems Lab, we perform research towards high-performance (embedded) digital systems spanning several topic areas, including machine learning, computer vision, and robotics.

MORE DETAILS

@CBouganis

Tweets by @CBouganis

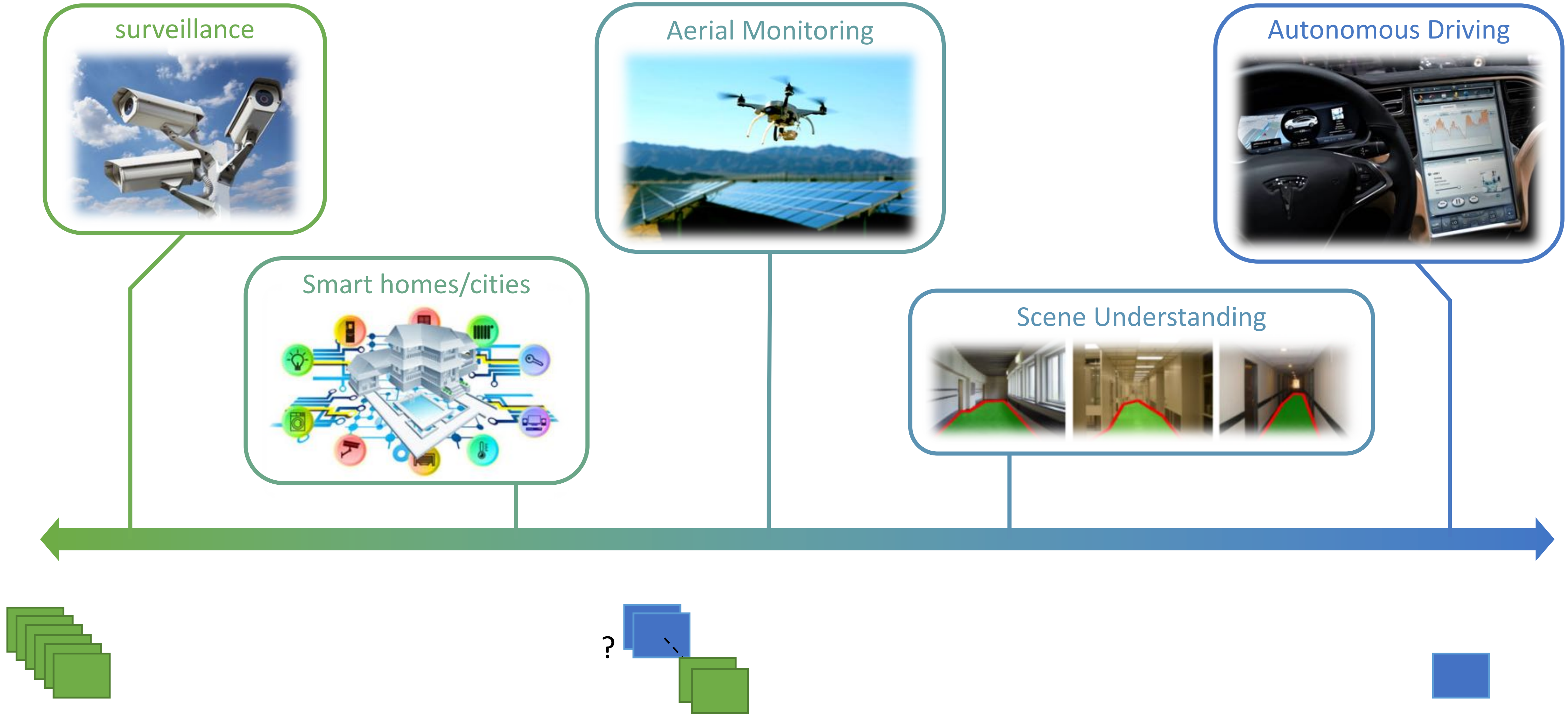
Christos
@CBouganis
We are recruiting for an exciting post on Machine Learning and FPGAs. Please see details here: tinyurl.com/yadhwend

Research Assistant/Associat...
The Intelligent Digital Systems ...
imperial.ac.uk

Aug 4, 2018

Christos
@CBouganis

DNNs in the Embedded Space – Variability in Performance Requirements



surveillance

Smart homes/cities

Aerial Monitoring

Scene Understanding

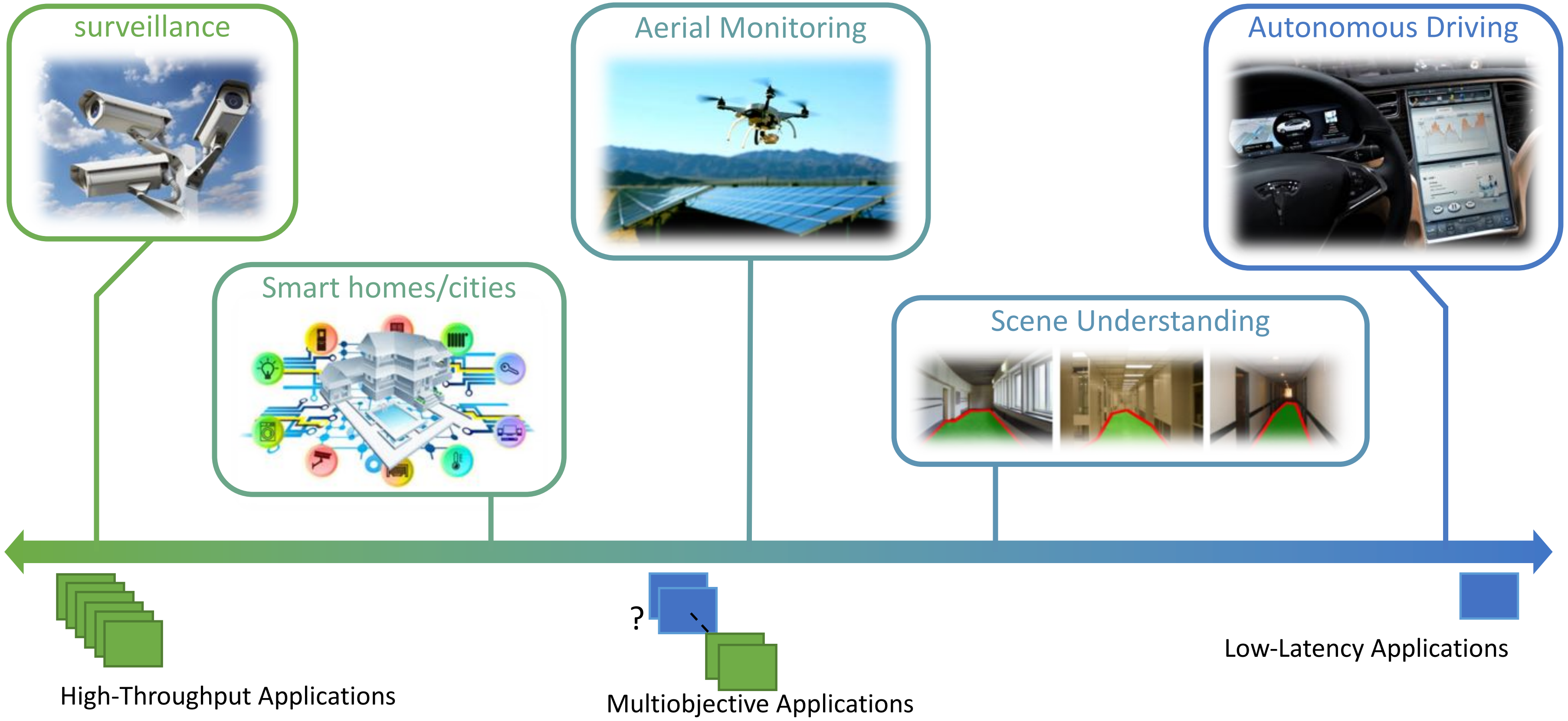
Autonomous Driving

High-Throughput Applications

Multiobjective Applications

Low-Latency Applications

DNNs in the Embedded Space – Variability in Performance Requirements



surveillance

Smart homes/cities

Aerial Monitoring

Scene Understanding

Autonomous Driving

High-Throughput Applications

Multiobjective Applications

Low-Latency Applications

Our approach: Couple the design of the ML algorithm with the design of the computational platform to improve performance and enable the deployment of AI systems

Power constraints

- Absolute power consumption
- Performance-per-Watt

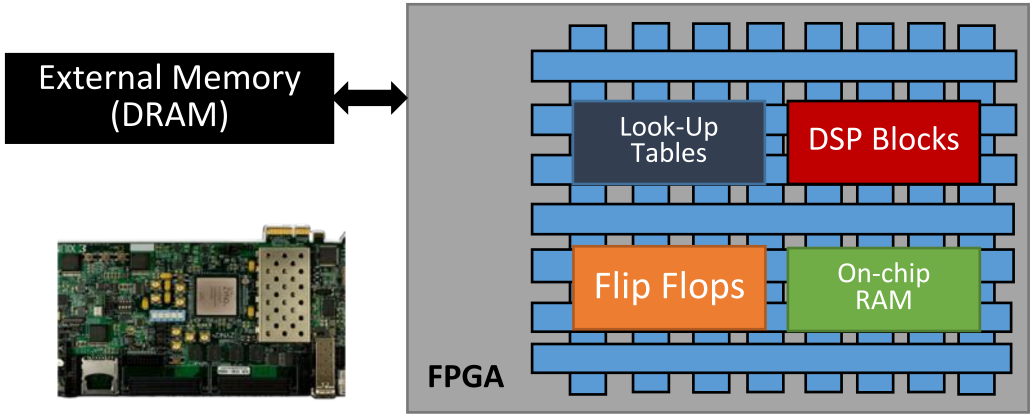
Conventional Embedded Platforms for Neural Networks

GPUs – Tegra K1, X1 and X2
DSPs – Qualcomm Hexagon,
Apple Neural Engine, …

✓ High throughput
✗ Low latency
✗ Low power
✓ Tools

FPGAs

- Custom datapath
- Custom memory subsystem
- Programmable interconnections
- Reconfigurability



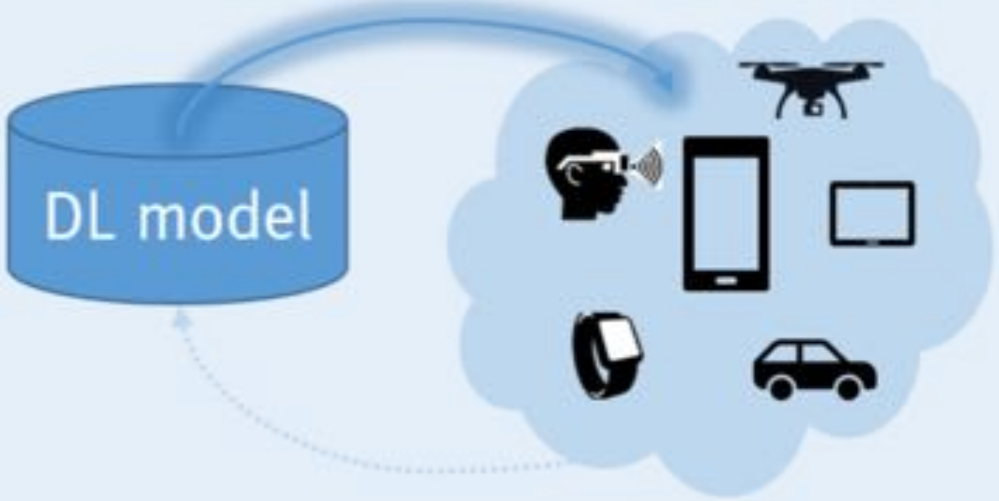
✓ High throughput
✓ Low latency
✓ Low power
✗ Tools

Challenge: Huge design space
Our Approach: Automated toolflows

Research Areas / Challenges

- Mapping Automation
- Multiple CNN Mapping
- Time-constrained Inference
- Privacy-aware Deep Learning

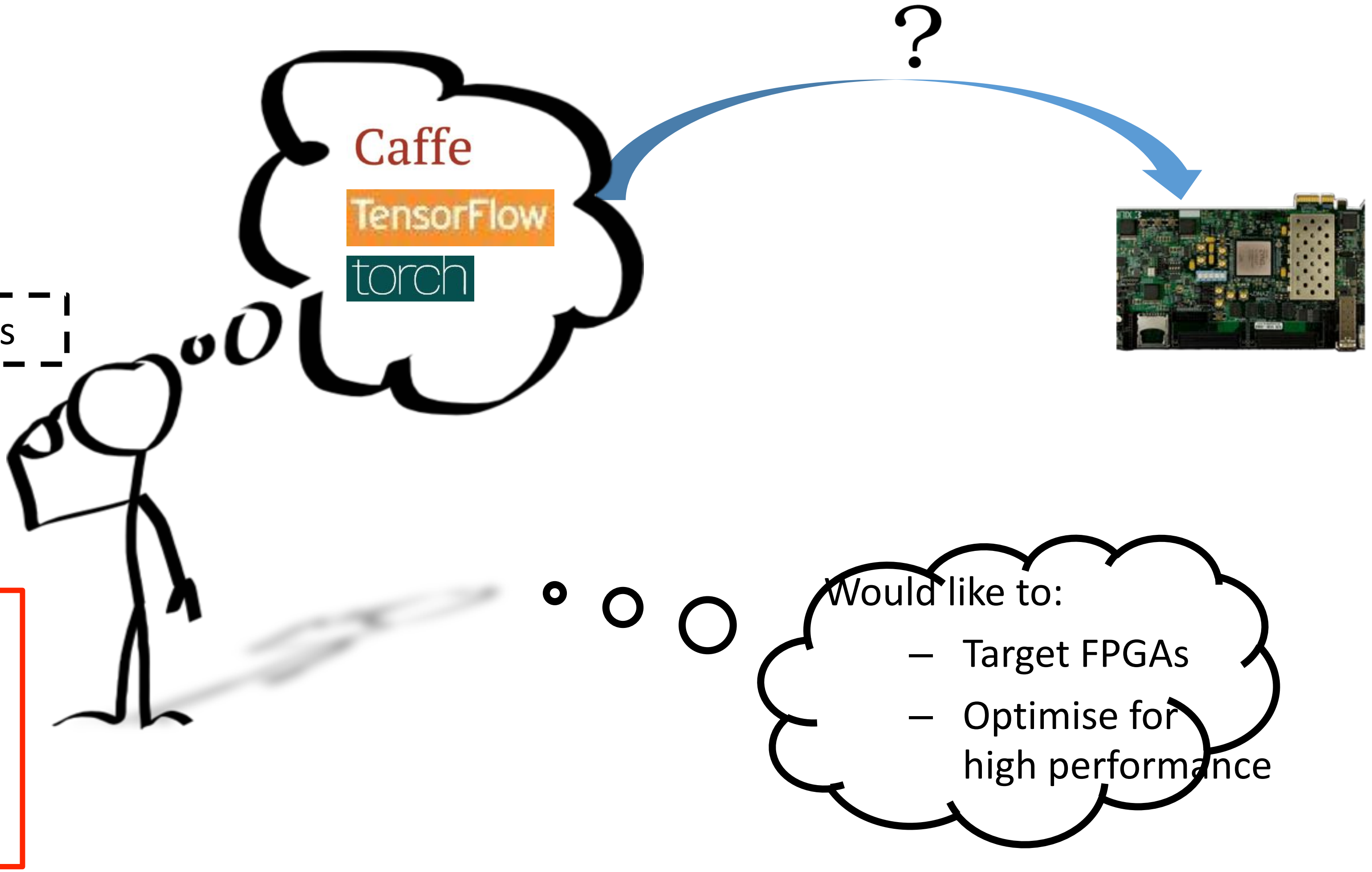
Challenge #1: Mapping Automation



Challenge #1: Mapping Automation

Little knowledge about FPGAs
Ease of deployment
“Good” designs

Deep Learning Developers



Would like to:

- Target FPGAs
- Optimise for high performance

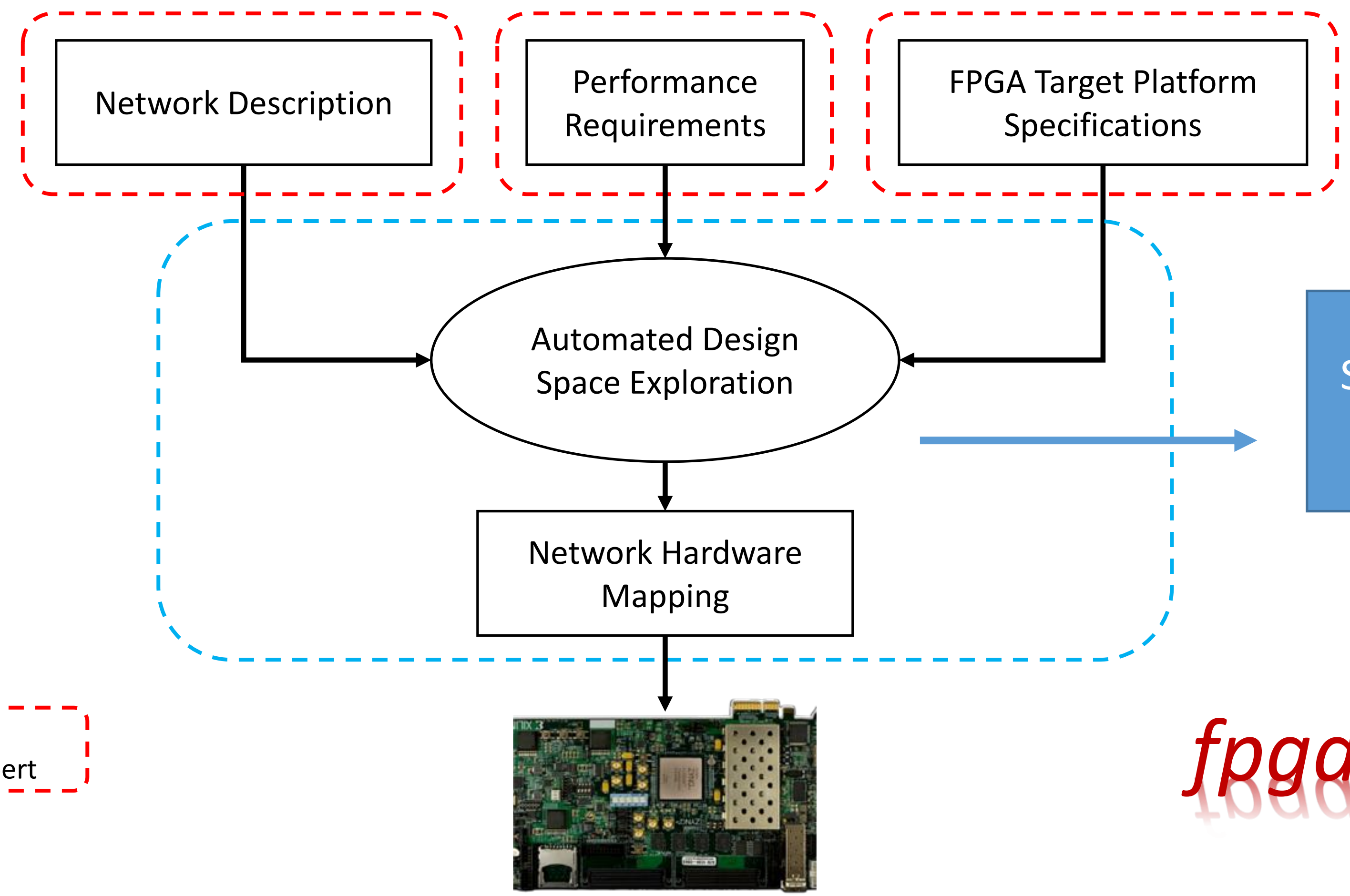
Challenges:

- High-dimensional design space
- Diverse application-level needs
- Utilise the FPGA resources
- Design automation

Challenge #1: Automated CNN-to-FPGA Toolflow

Caffe

torch


Network Description
Performance Requirements
FPGA Target Platform Specifications
Automated Design Space Exploration
Network Hardware Mapping

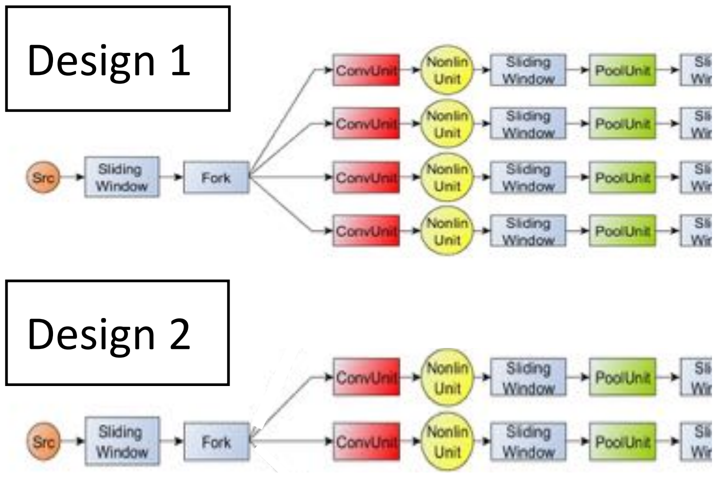
SDFG : allows principled design space exploration

Supplied by
Deep Learning Expert

fpgaConvNet

fpgaConvNet – Design Space Exploration and Optimisation

- Synchronous Dataflow Modelling
  - Capture hardware mappings as matrices
  - Transformations as *algebraic operations*
  - Analytical *performance model*
  - Cast design space exploration as a mathematical optimisation problem



Hardware Stages

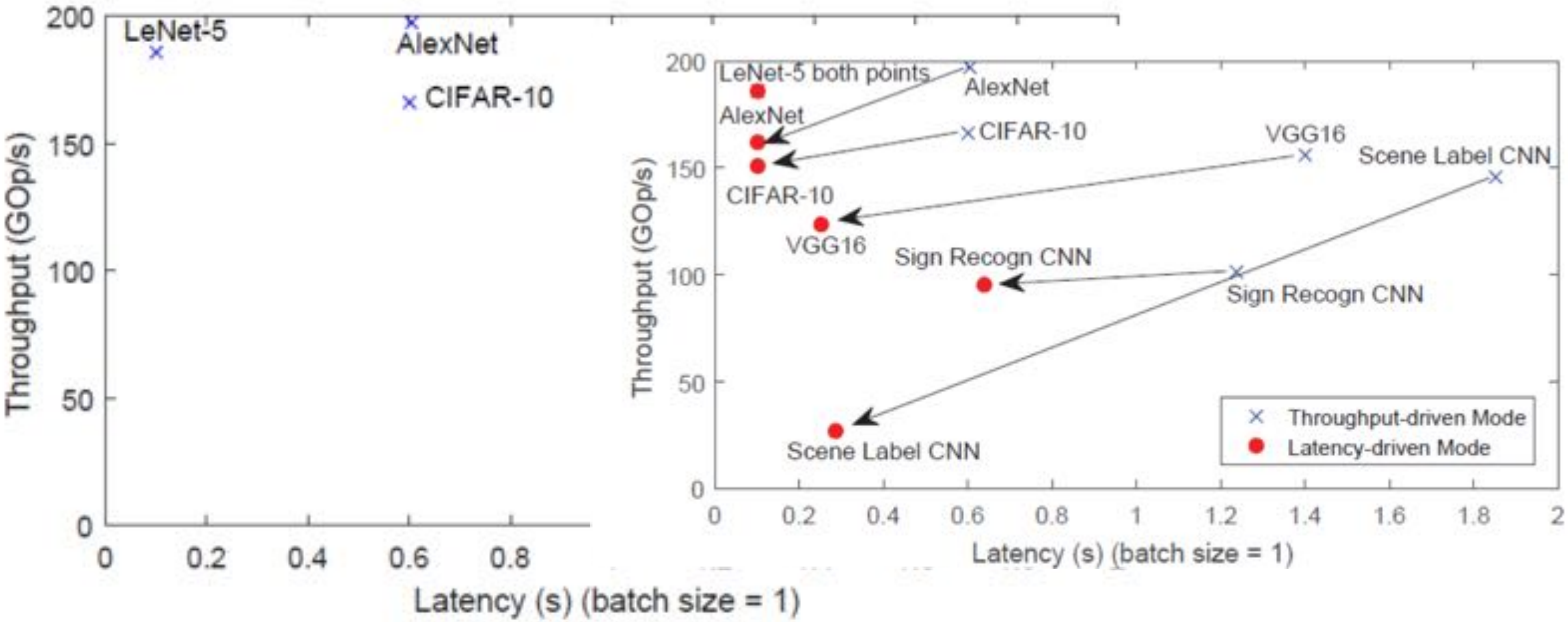
Interconnections

$$\Gamma_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K^2 & -K^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4K^2 & -4K^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4P^2 & -4P^2 \end{bmatrix}$$

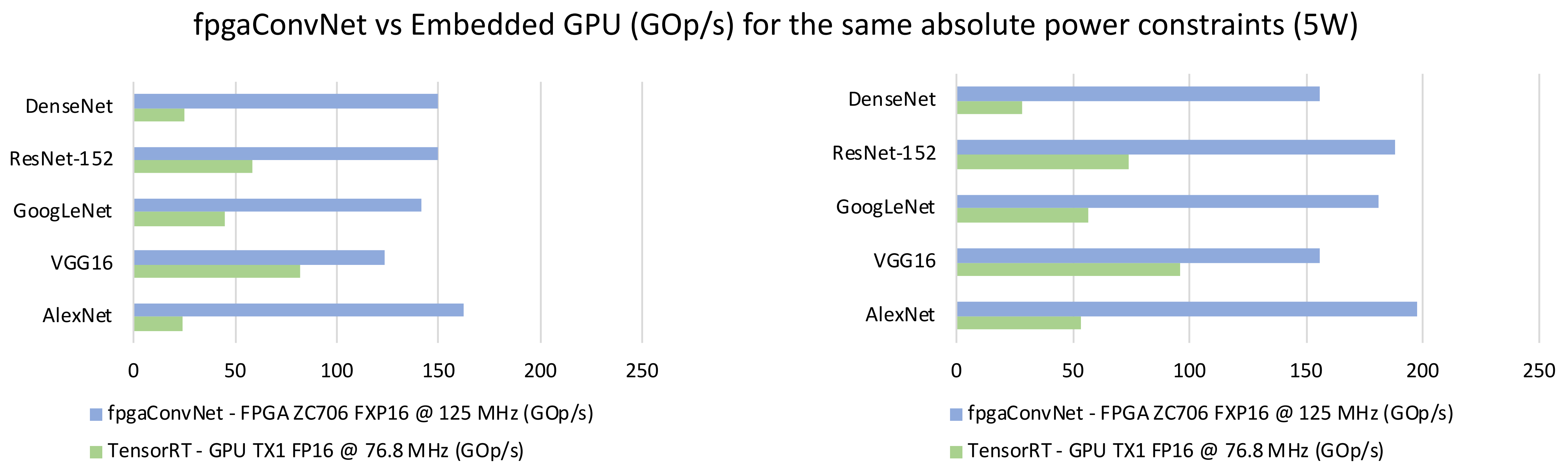
$$\Gamma_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K^2 & -K^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2K^2 & -2K^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2P^2 & -2P^2 \end{bmatrix}$$

$$t_{total}(B, N_P, \mathbf{\Gamma}) = \sum_{i=1}^{N_P} t_i(B, \mathbf{\Gamma}_i) + (N_P - 1) \cdot t_{reconfig.}$$

Meeting the performance requirements



Comparison with Embedded GPUs: Same absolute power constraints (5W)

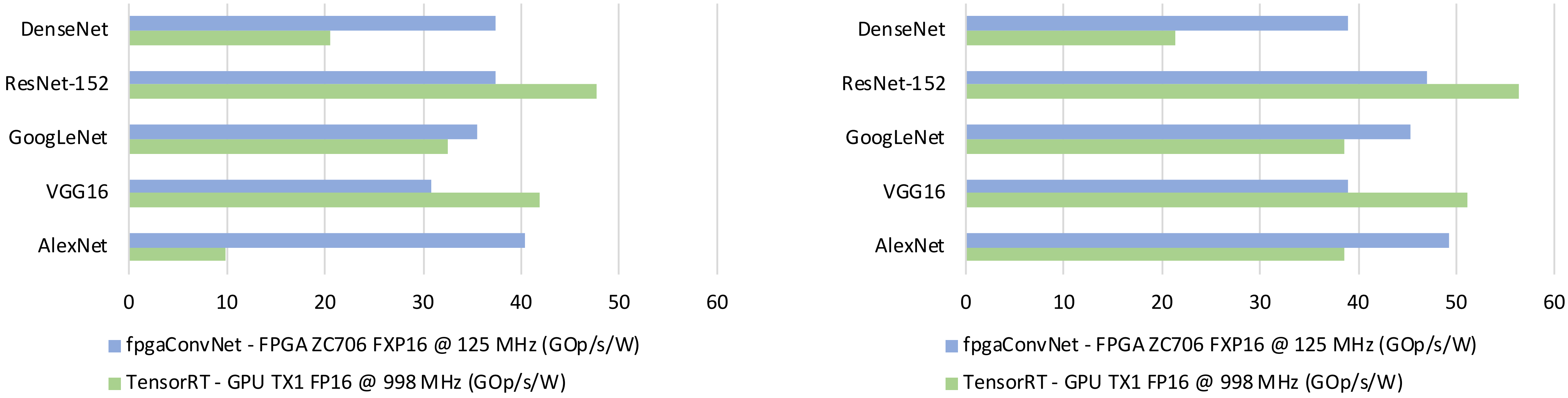

fpgaConvNet vs Embedded GPU (GOp/s) for the same absolute power constraints (5W)

- Latency-driven scenario → batch size of 1
- Up to 6.65× speedup with an average of 3.95× (3.43× geo. mean)

- Throughput-driven scenario → favourable batch size
- Up to 5.53× speedup with an average of 3.32× (3.07× geo. mean)

Comparison with Embedded GPUs: Performance-per-Watt

fpgaConvNet vs Embedded GPU (GOp/s/W)

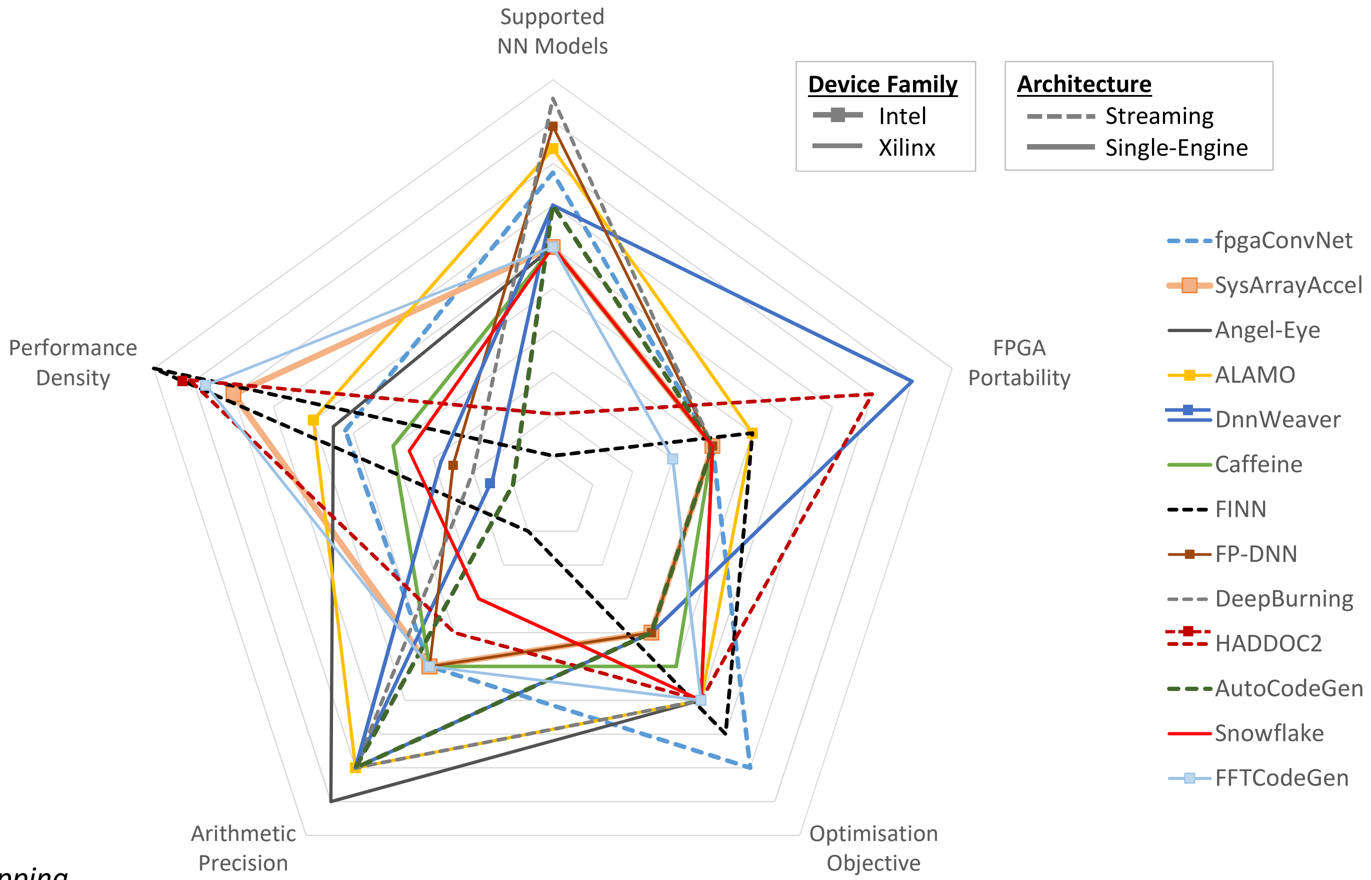


- Latency-driven scenario → batch size of 1
- Average of 1.70× (1.36× geo. mean) in GOp/s/W

- Throughput-driven scenario → favourable batch size
- Average of 1.17× (1.12× geo. mean) in GOp/s/W

Other approaches

Toolflow Name	Interface	Year
fpgaConvNet [86][87][88][85]	Caffe & Torch	May 2016
DeepBurning [90]	Caffe	June 2016
Angel-Eye [68][23][24]	Caffe	July 2016
ALAMO [58][56][57][55][59]	Caffe	August 2016
Haddoc2 [1][2]	Caffe	September 2016
DnnWeaver [75][76]	Caffe	October 2016
Caffeine [98]	Caffe	November 2016
AutoCodeGen [54]	Proprietary Input Format	December 2016
Finn [84][19]	Theano	February 2017
FP-DNN [22]	TensorFlow	May 2017
Snowflake [21][10]	Torch	May 2017
SysArrayAccel [91]	C Program	June 2017
FFTCodeGen [100][97][96][95]	Proprietary Input Format	December 2017

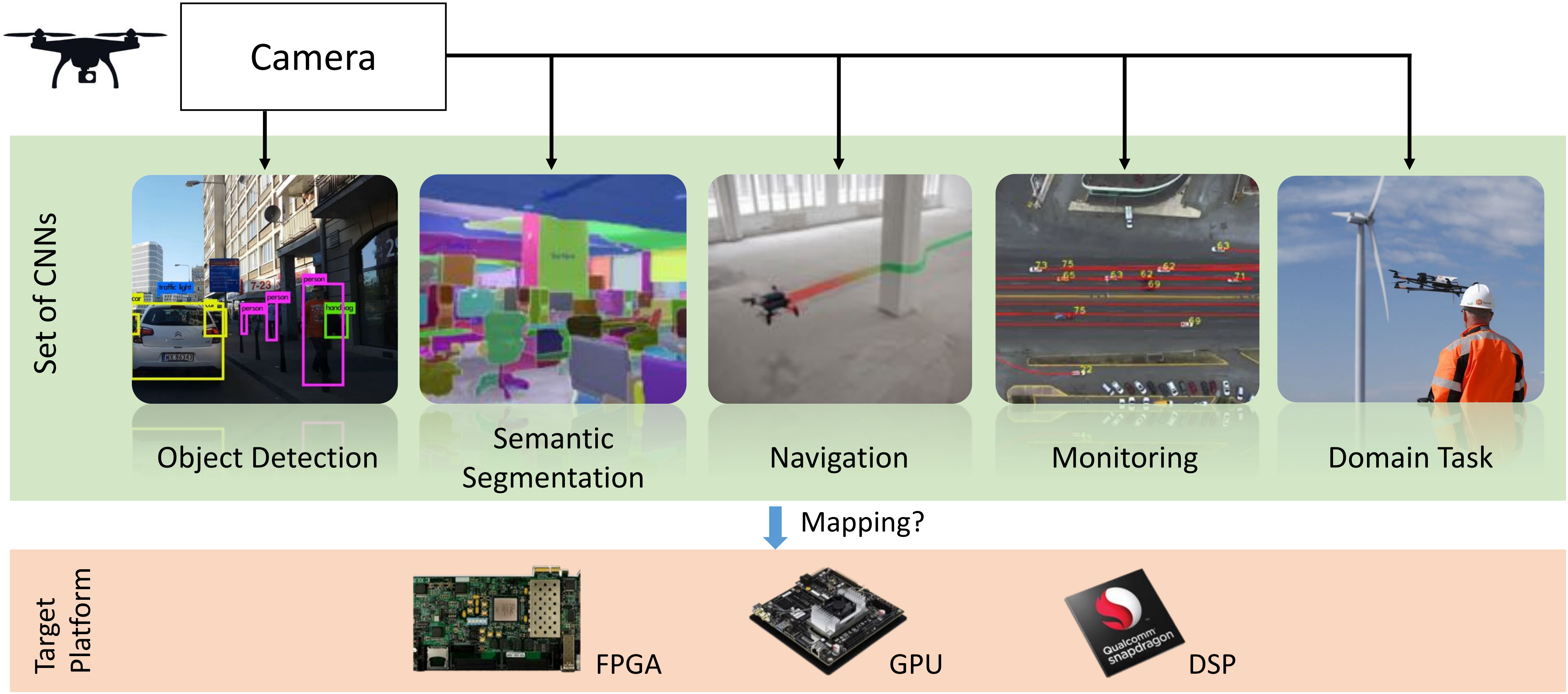


Stylianos I. Venieris, Alexandros Kouris and Christos-Savvas Bouganis, "*Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions*", ACM Computing Surveys, 2018

Imperial College London

Challenge #2: Multi-CNN Systems

Challenge #2: Multi-CNN Systems – Autonomous Drones



Camera

Set of CNNs

Object Detection

Semantic Segmentation

Navigation

Monitoring

Domain Task

Mapping?

Target Platform

FPGA

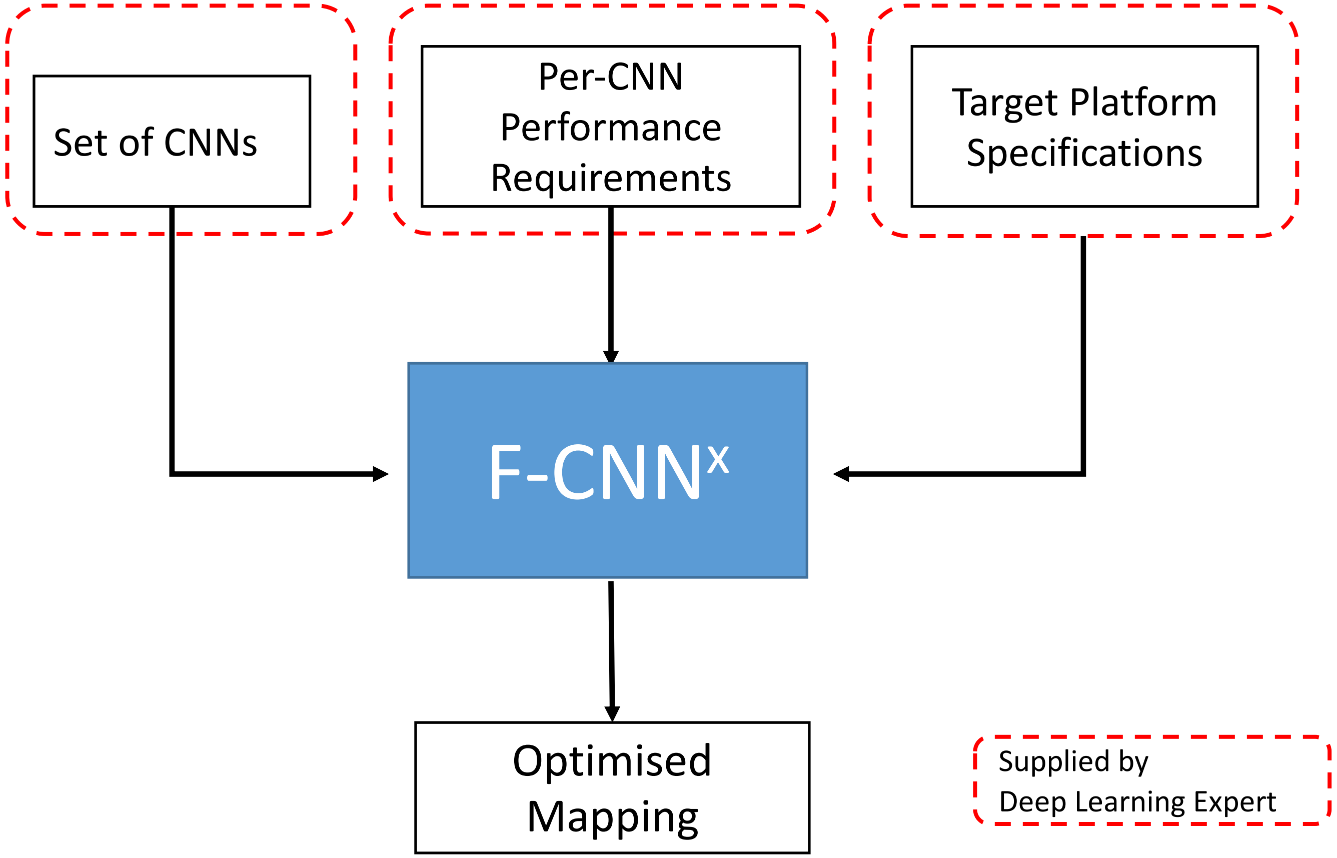
GPU

DSP

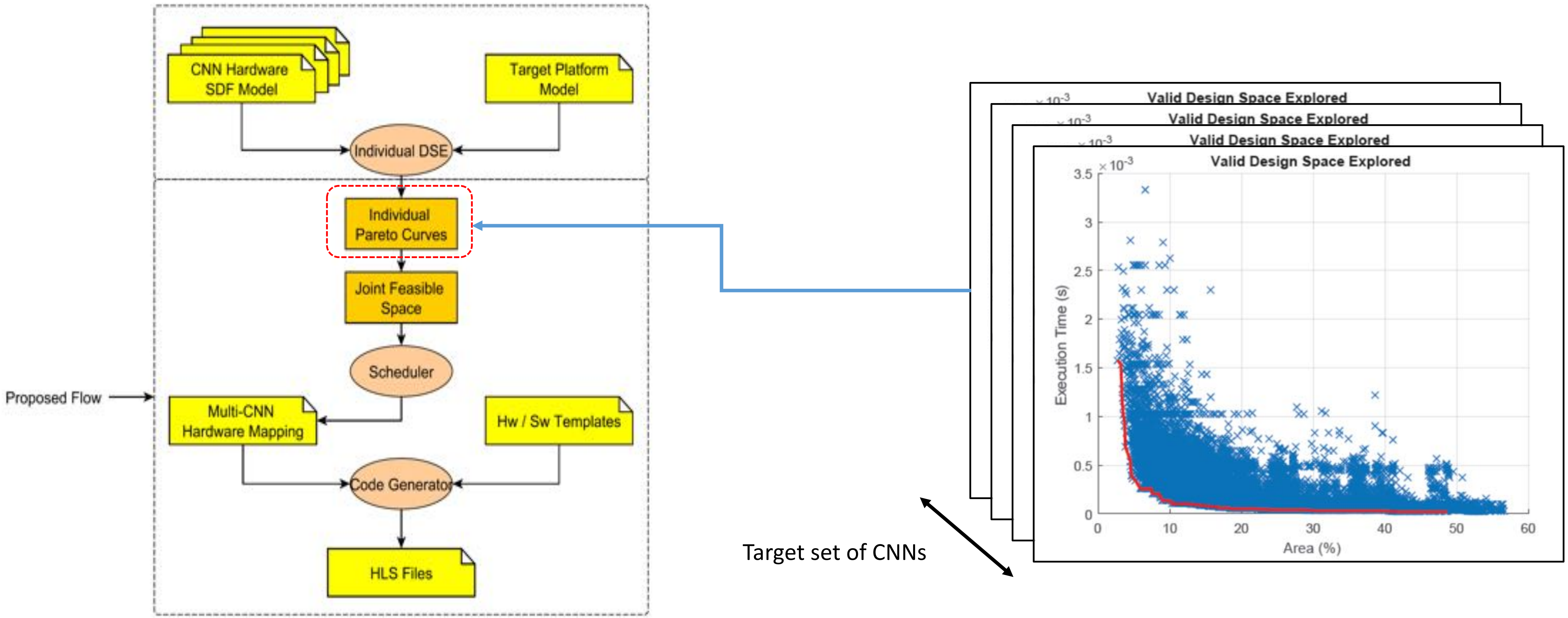
Challenge #2: Multi-DNN System

Challenges:

- Resource allocation among CNNs
- Design automation
- Models with different performance constraints, e.g. required throughput and latency
- Competing for the same pool of resources
- High-dimensional design space

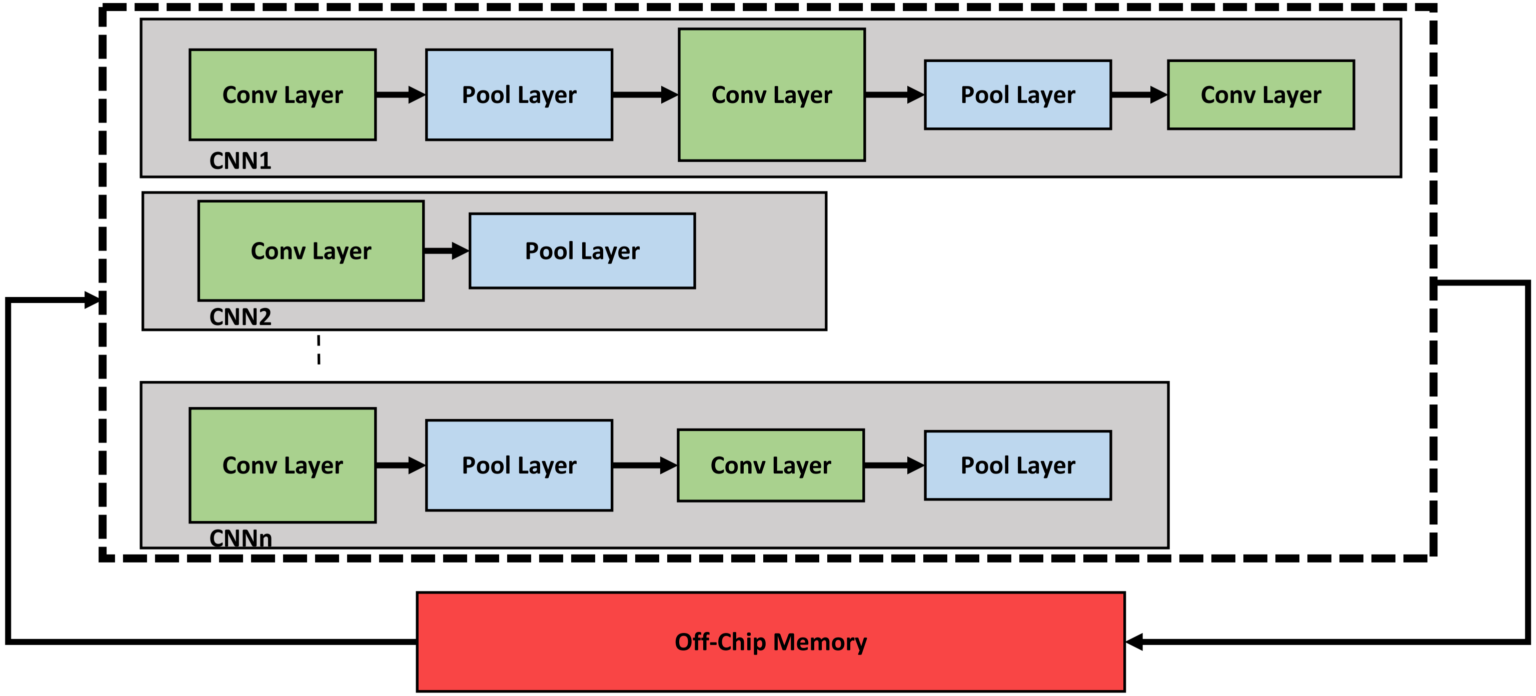

Set of CNNs
Per-CNN Performance Requirements
Target Platform Specifications
F-CNN^x
Optimised Mapping
Supplied by Deep Learning Expert

Proposed Design Space Exploration Method

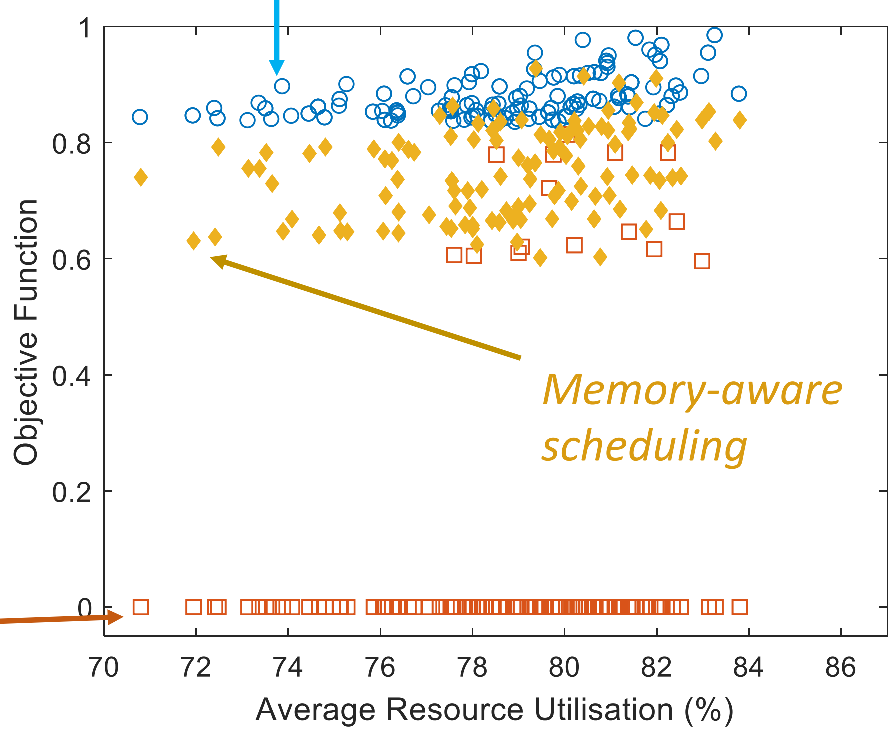


Imperial College London

FPGA Architecture



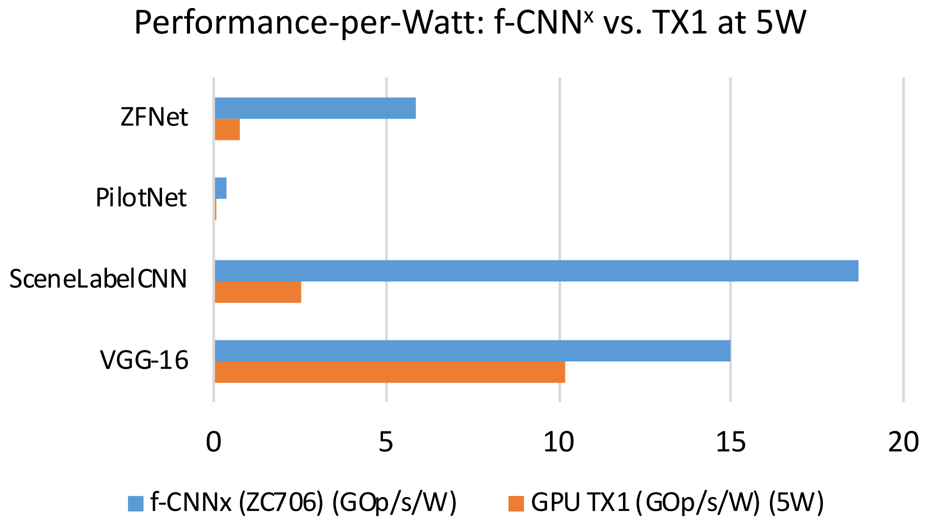
Full platform available bandwidth for each CNN engine

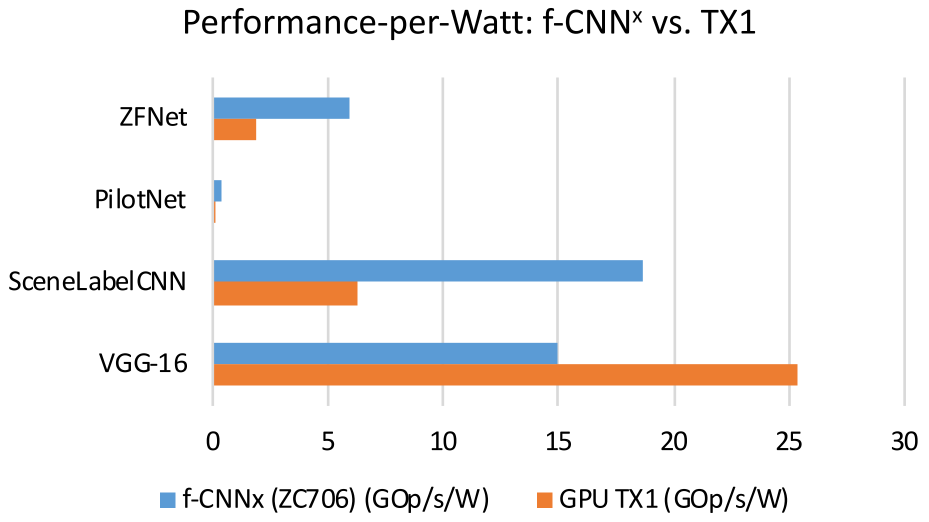


Memory-aware scheduling

Memory-unaware scheduling

Comparison with Embedded GPUs



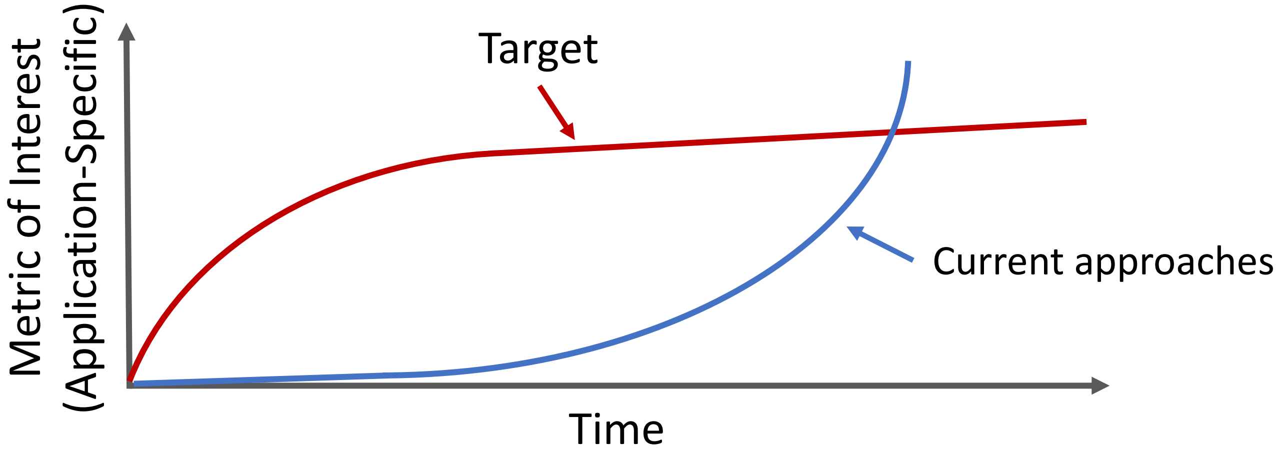


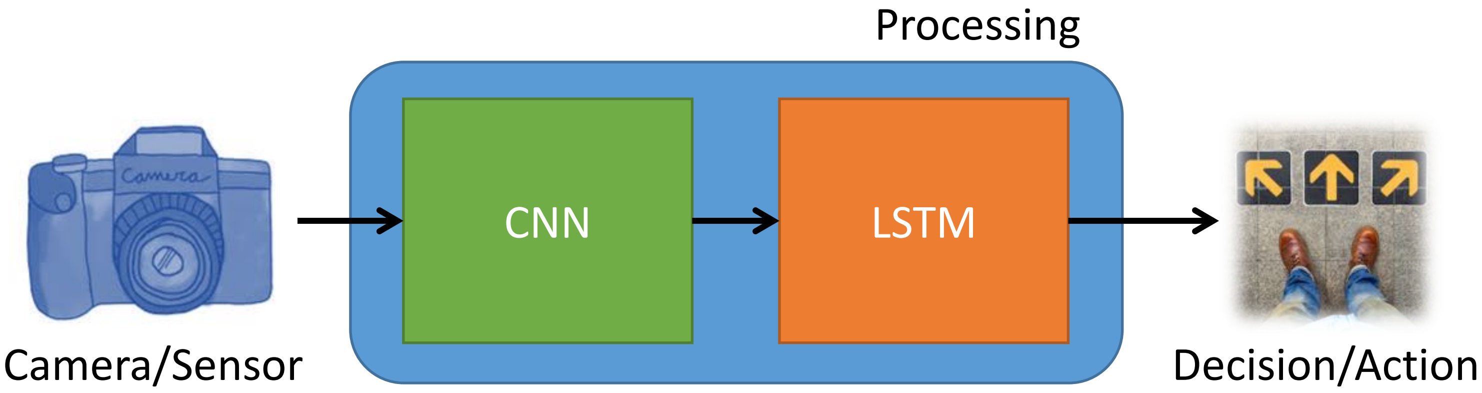
- Latency-driven scenario → batch size of 1
- Up to 19.09× speedup with an average of 6.85× (geo. mean)

- Latency-driven scenario → batch size of 1
- Up to 9.61× speedup with an average of 2.76× (geo. mean)

Challenge #3: Time-constrained Inference

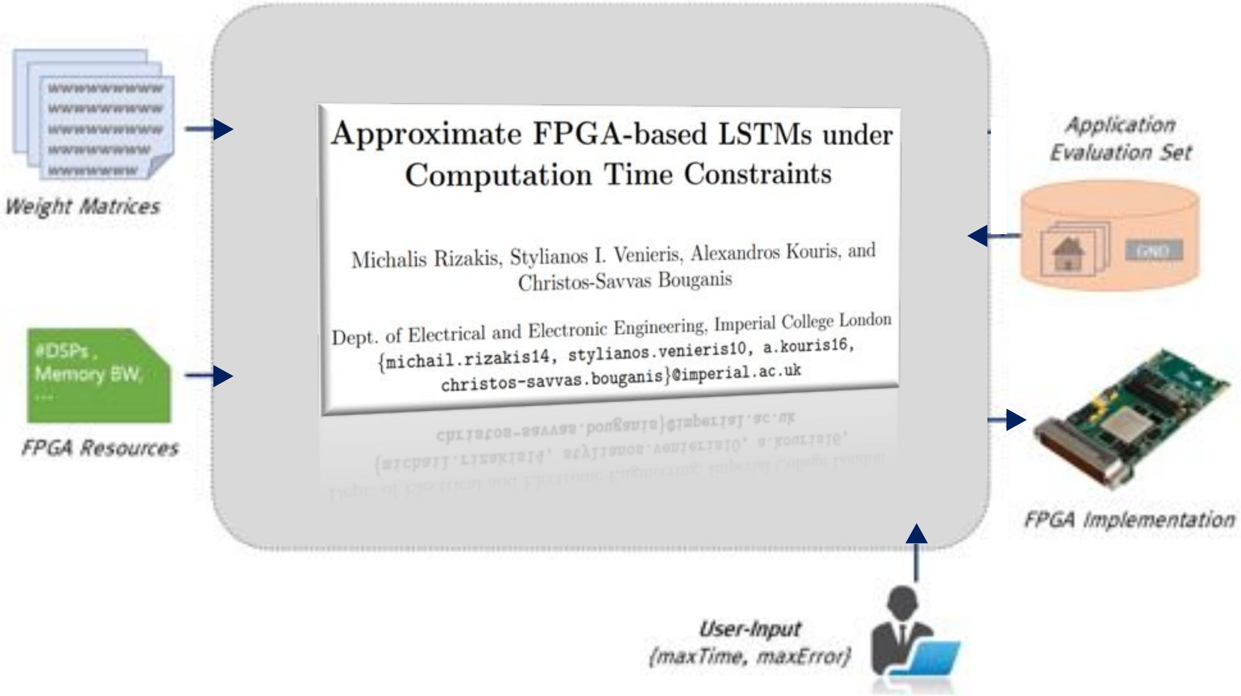
Challenge #3: Time-constrained Inference





Challenge #3: Time-constrained Inference

- Approximate LSTMs
  - Iterative refinement using SVD + Pruning.
  - Parametrized with respect to:
    - Number of iterations
    - Level of pruning
- Parametrized hardware architecture, tailored for approximate LSTMs
- Co-optimise given a user-defined time budget



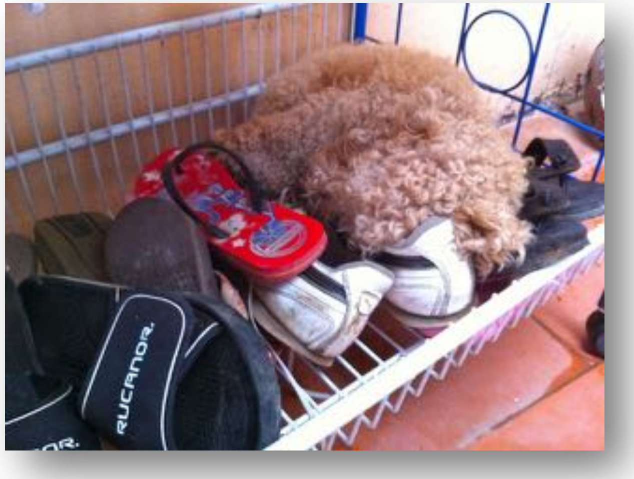
Approximate FPGA-based LSTMs under Computation Time Constraints

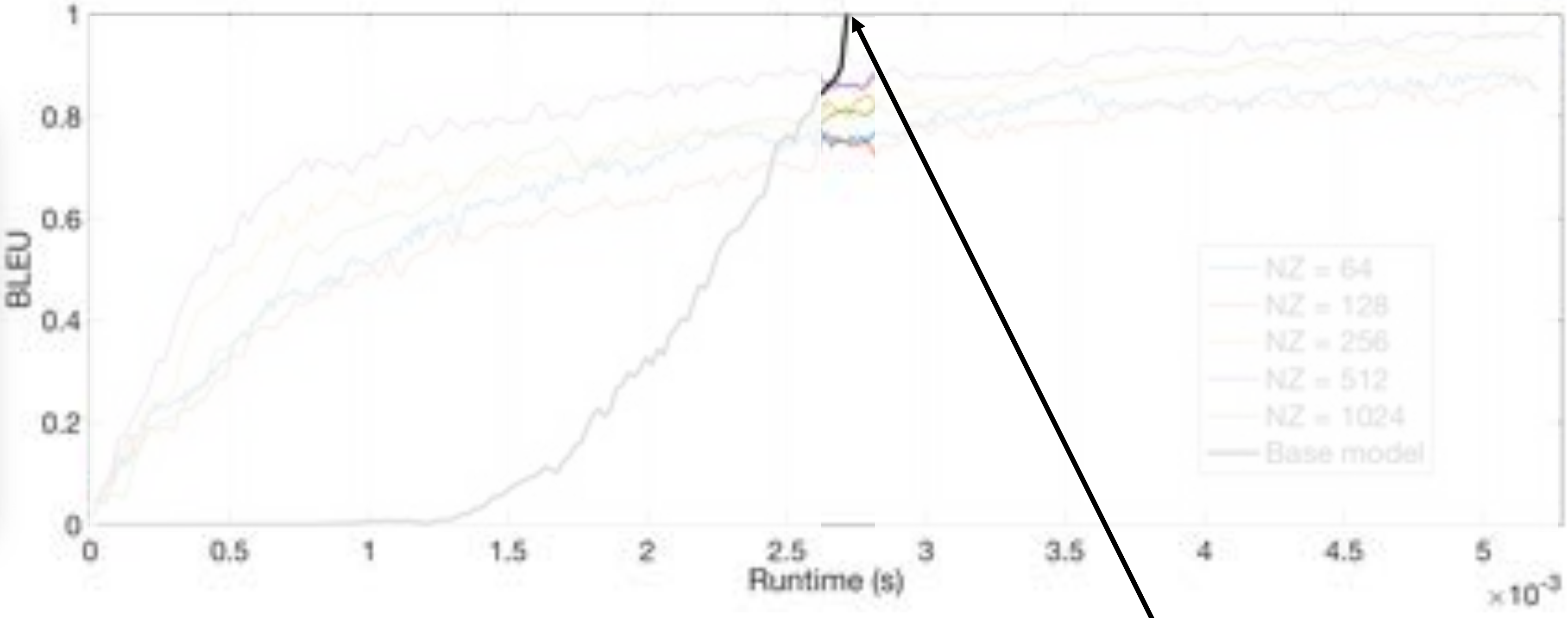
Michalis Rizakis, Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis

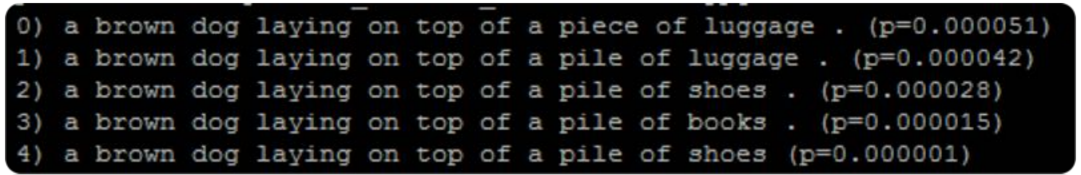
Dept. of Electrical and Electronic Engineering, Imperial College London
{michail.rizakis14, stylianos.venieris10, a.kouris16, christos-savvas.bouganis}@imperial.ac.uk

Impact on LSTM-based Image Captioning

Input Image



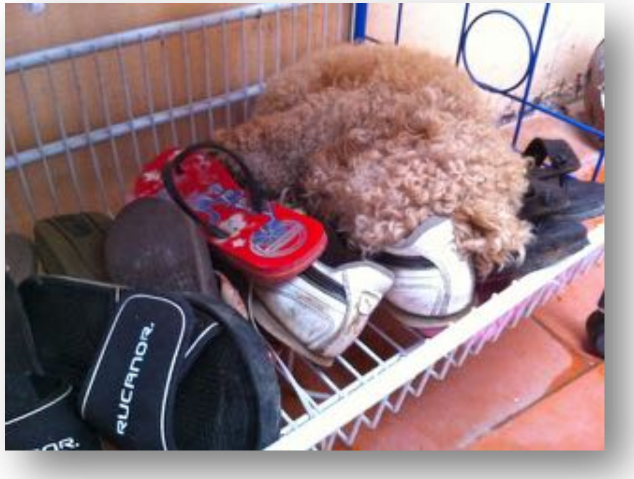


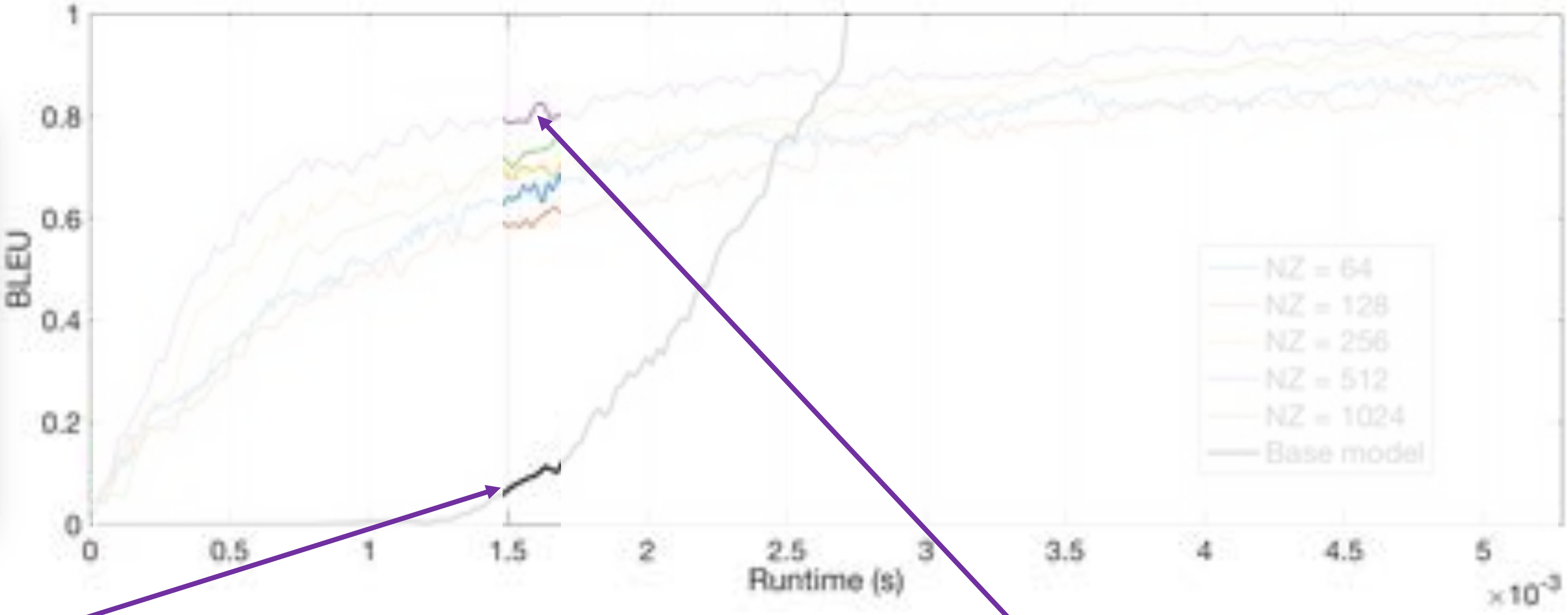


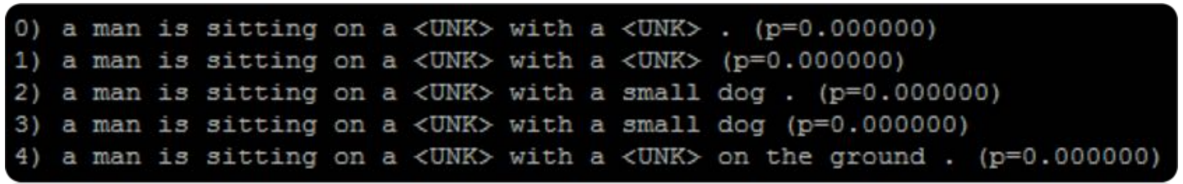
```
0) a brown dog laying on top of a piece of luggage . (p=0.000051)
1) a brown dog laying on top of a pile of luggage . (p=0.000042)
2) a brown dog laying on top of a pile of shoes . (p=0.000028)
3) a brown dog laying on top of a pile of books . (p=0.000015)
4) a brown dog laying on top of a pile of shoes (p=0.000001)
```

Impact on LSTM-based Image Captioning

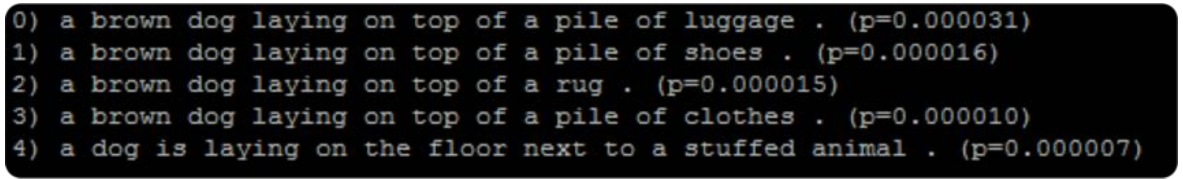
Input Image







```
0) a man is sitting on a <UNK> with a <UNK> . (p=0.000000)
1) a man is sitting on a <UNK> with a <UNK> (p=0.000000)
2) a man is sitting on a <UNK> with a small dog . (p=0.000000)
3) a man is sitting on a <UNK> with a small dog (p=0.000000)
4) a man is sitting on a <UNK> with a <UNK> on the ground . (p=0.000000)
```



```
0) a brown dog laying on top of a pile of luggage . (p=0.000031)
1) a brown dog laying on top of a pile of shoes . (p=0.000016)
2) a brown dog laying on top of a rug . (p=0.000015)
3) a brown dog laying on top of a pile of clothes . (p=0.000010)
4) a dog is laying on the floor next to a stuffed animal . (p=0.000007)
```

Challenge #4: Privacy-aware Deep Learning

Challenge #4: Privacy-restricted Optimisation

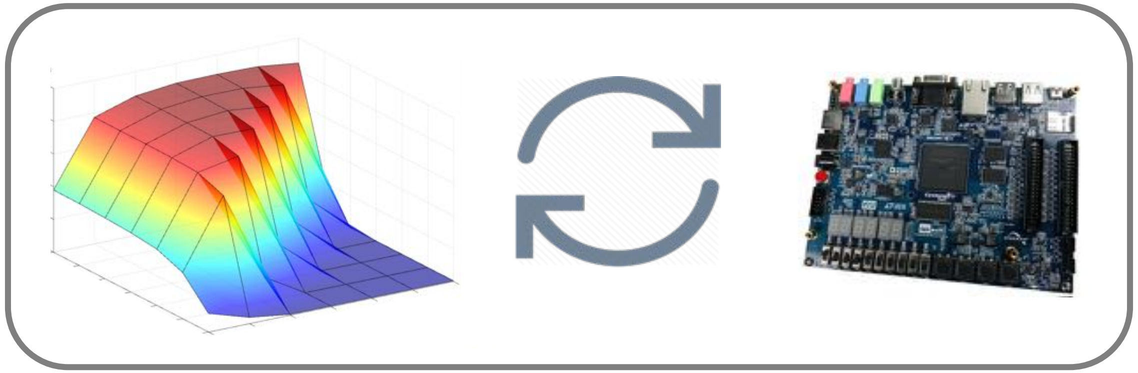
Aim: Design an optimised HW system (performance and accuracy)

Given:

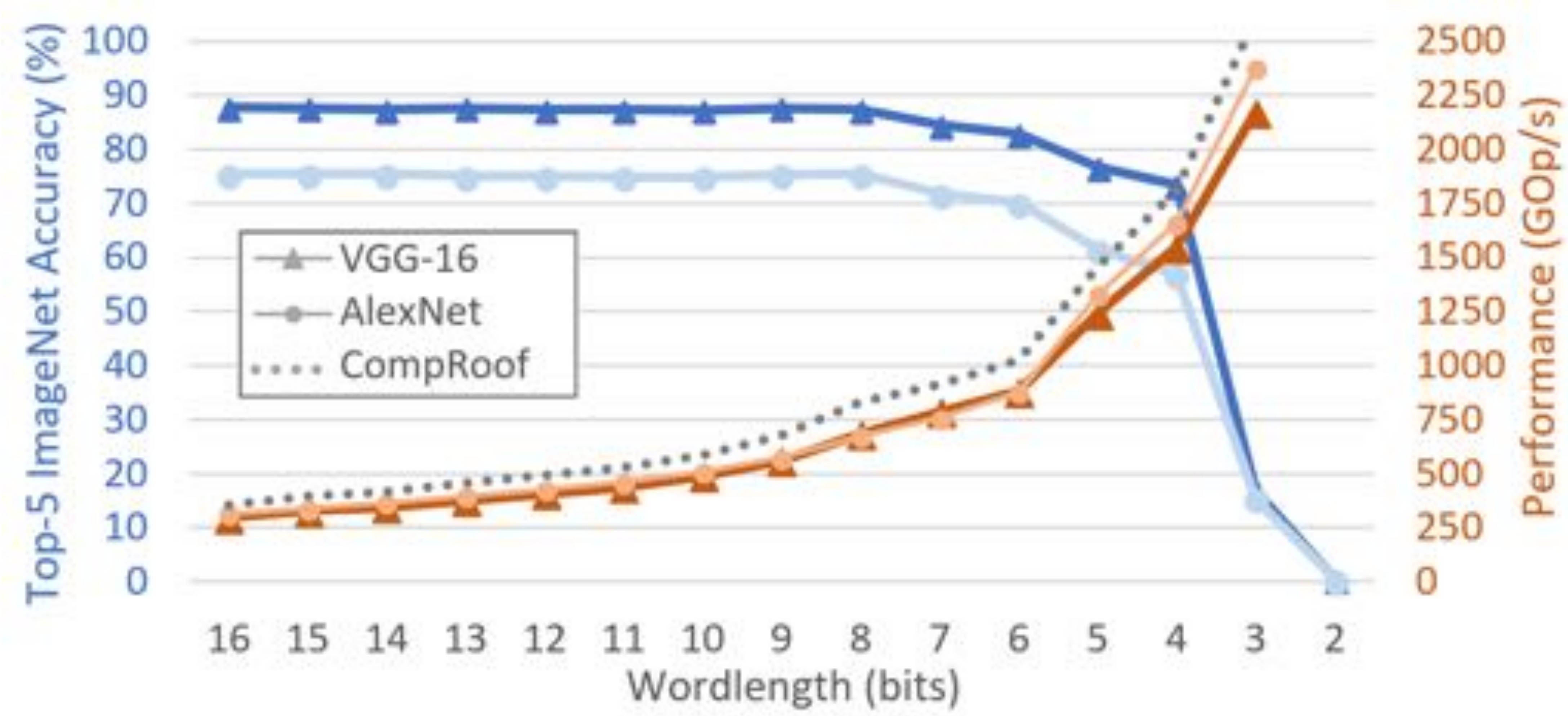
- A High-Level CNN Description (i.e. Caffe)
- A target FPGA platform
- ***Traini✗Data*** *privacy, availability*
- ***Testing Data***
- **Target metric (top1/top-5 accuracy, …)**

➔ ***quantisation with retraining step***

Limited quantisation opportunities

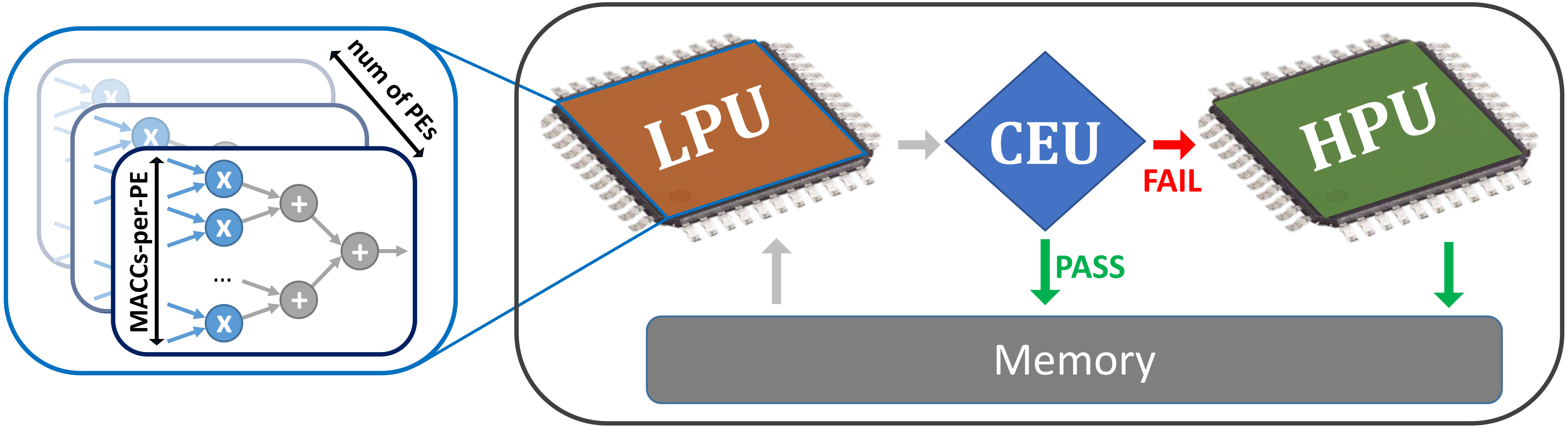


Challenge #4: Privacy-aware Deep Learning



CascadeCNN: High-Level System Architecture

- Pushing quantization bellow limits of acceptable accuracy to gain performance (high throughput)
- Evaluation of Quality of Prediction to identify and correct error introduced by quantization



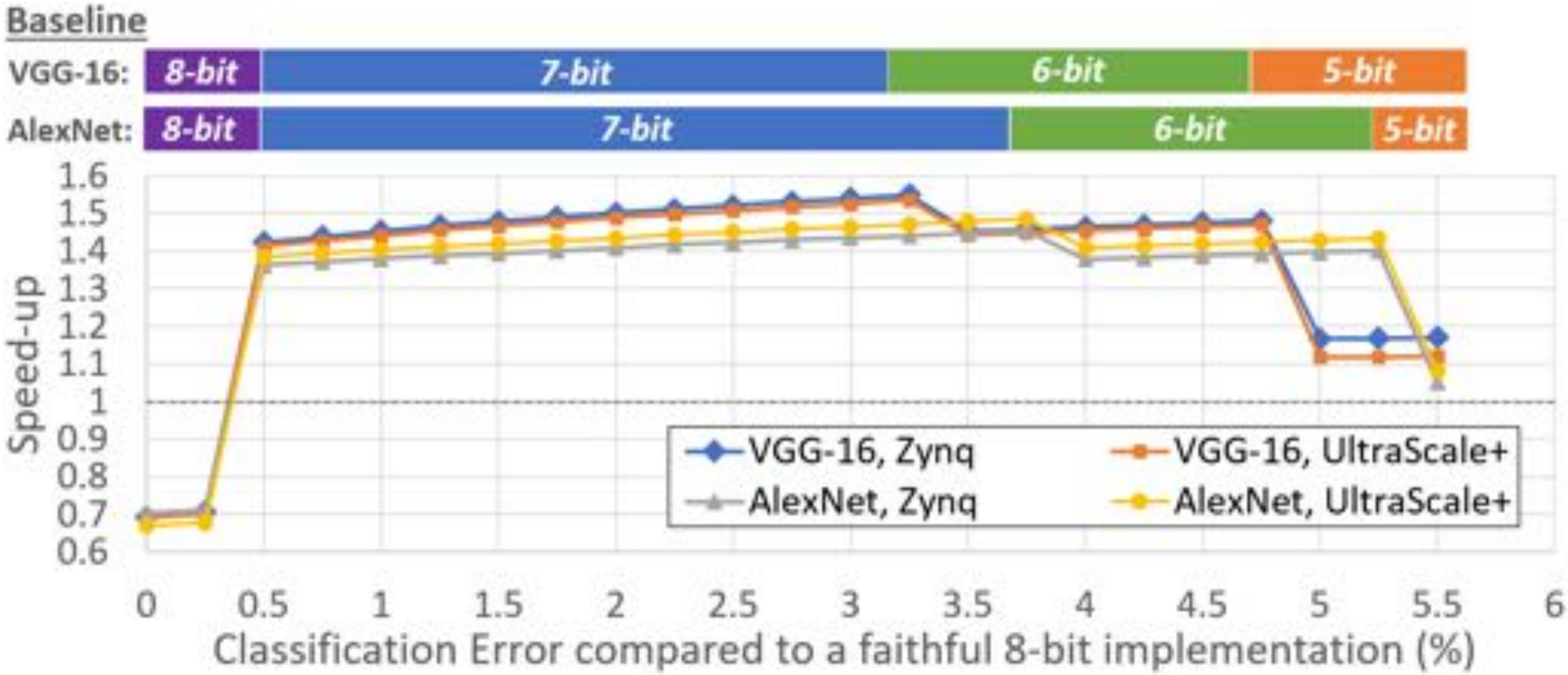
Low-Precision Unit: Degraded accuracy classification with high performance

Confidence Evaluation Unit: Identify misclassified cases

High-Precision Unit: Correct detected misclassified samples, to restore accuracy

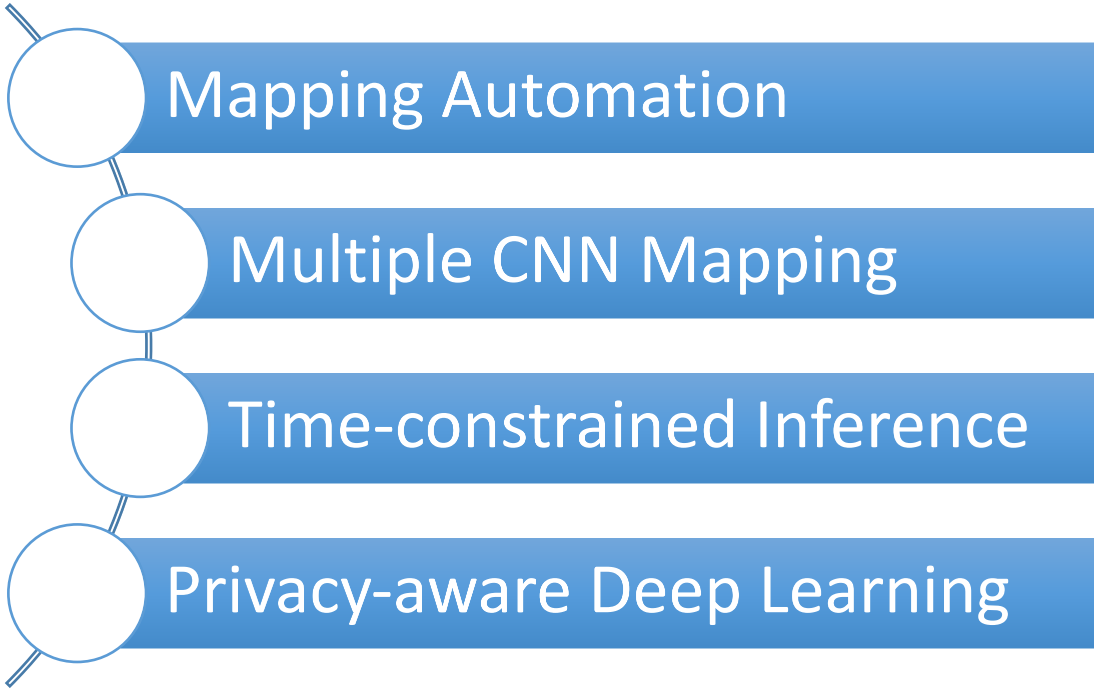
Imperial College London

$Cascade^{C}{}_{N_N}$: Results



Summary

Research topics



- Mapping Automation
- Multiple CNN Mapping
- Time-constrained Inference
- Privacy-aware Deep Learning

A. Kouris and C-S Bouganis, "Learning to Fly by MySelf: A Self-Supervised CNN-based Approach for Autonomous Navigation", IROS, 2018

Publications

- Alexandros Kouris, Stylianos I. Venieris, and Christos-Savvas Bouganis. 2018. ***CascadeCNN: Pushing the performance limits of quantisation.*** *In SysML.*
- Alexandros Kouris, Stylianos I. Venieris, and Christos-Savvas Bouganis. 2018. ***CascadeCNN: Pushing the Performance Limits of Quantisation in Convolutional Neural Networks***. *In 2018 28th International Conference on Field Programmable Logic and Applications (FPL).*
- C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C. S. Bouganis. 2018. ***DroNet: Efficient Convolutional Neural Network Detector for Real-Time UAV Applications.*** *In 2018 Design, Automation Test in Europe Conference Exhibition (DATE). 967–972.*
- Michalis Rizakis, Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. 2018. ***Approximate FPGA-based LSTMs under Computation Time Constraints.*** *In Applied Reconfigurable Computing - 14th International Symposium, ARC 2018, Santorini, Greece, May 2 - 4, 2018, 3–15.*
- Stylianos I. Venieris and Christos-Savvas Bouganis. 2016. ***fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs.*** *In 2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). 40–47.*
- Stylianos I. Venieris and Christos-Savvas Bouganis. 2017. ***fpgaConvNet: A Toolflow for Mapping Diverse Convolutional Neural Networks on Embedded FPGAs.*** *In NIPS 2017 Workshop on Machine Learning on the Phone and other Consumer Devices.*
- Stylianos I. Venieris and Christos-Savvas Bouganis. 2017. ***fpgaConvNet: Automated Mapping of Convolutional Neural Networks on FPGAs*** (Abstract Only). *In Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 291–292.*
- S. I. Venieris and C. S. Bouganis. 2017. ***Latency-Driven Design for FPGA-based Convolutional Neural Networks.*** *In 2017 27th International Conference on Field Programmable Logic and Applications (FPL).*
- S. I. Venieris and C. S. Bouganis. 2018. ***f-CNNx: A Toolflow for Mapping Multiple Convolutional Neural Networks on FPGAs.*** *In 2018 28th International Conference on Field Programmable Logic and Applications (FPL).*
- Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. 2018. ***Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions***. *In ACM Computing Surveys 51, 3, Article 56 (June 2018), 39 pages.*