Deploying Deep Neural Networks in the Embedded Space

Stylianos I. Venieris, Alexandros Kouris, Christos-Savvas Bouganis

2nd International Workshop on
Embedded and Mobile Deep Learning (***EMDL***)

MobiSys, 15 June 2018

intelligent Digital Systems Lab
Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

Who we are

Stylianos I. Venieris
Machine Learning

Alexandros Kouris
Machine Learning,
Robotics

Konstantinos Boikos
Computer Vision,
SLAM

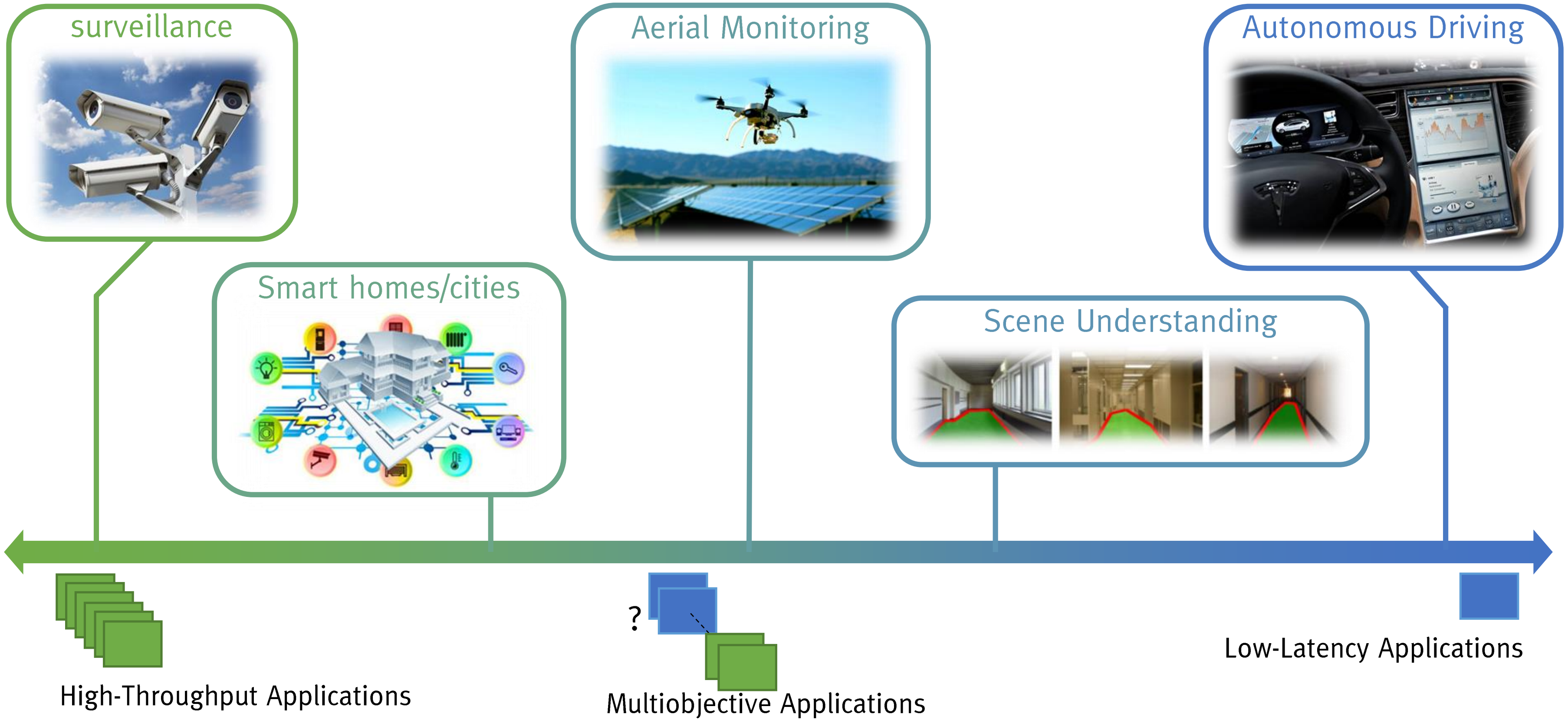
Manolis Vasileiadis
Computer Vision

Mudhar Bin Rabieah
Machine Learning

Nur Ahmadi
Brain-Machine Interface

Christos-Savvas Bouganis
Lab Director
Reader at
Imperial College London

DNNs in the Embedded Space – Variability in Performance Requirements



Focus: 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, …

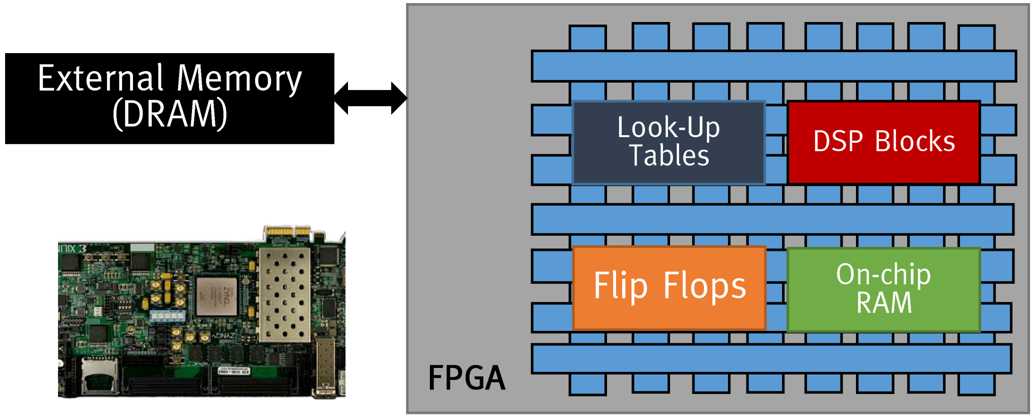
✓ High throughput

✗ Low latency

✗ Low power

FPGAs

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



✓ High throughput

✓ Low latency

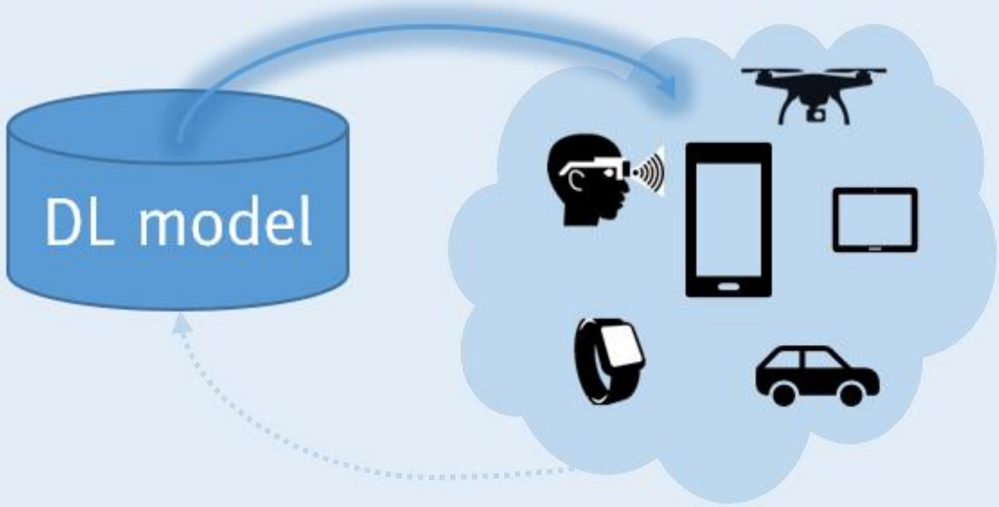
✓ Low power

Challenge: Huge design space
Our Approach: Automated toolflows

Research Areas / Challenges

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

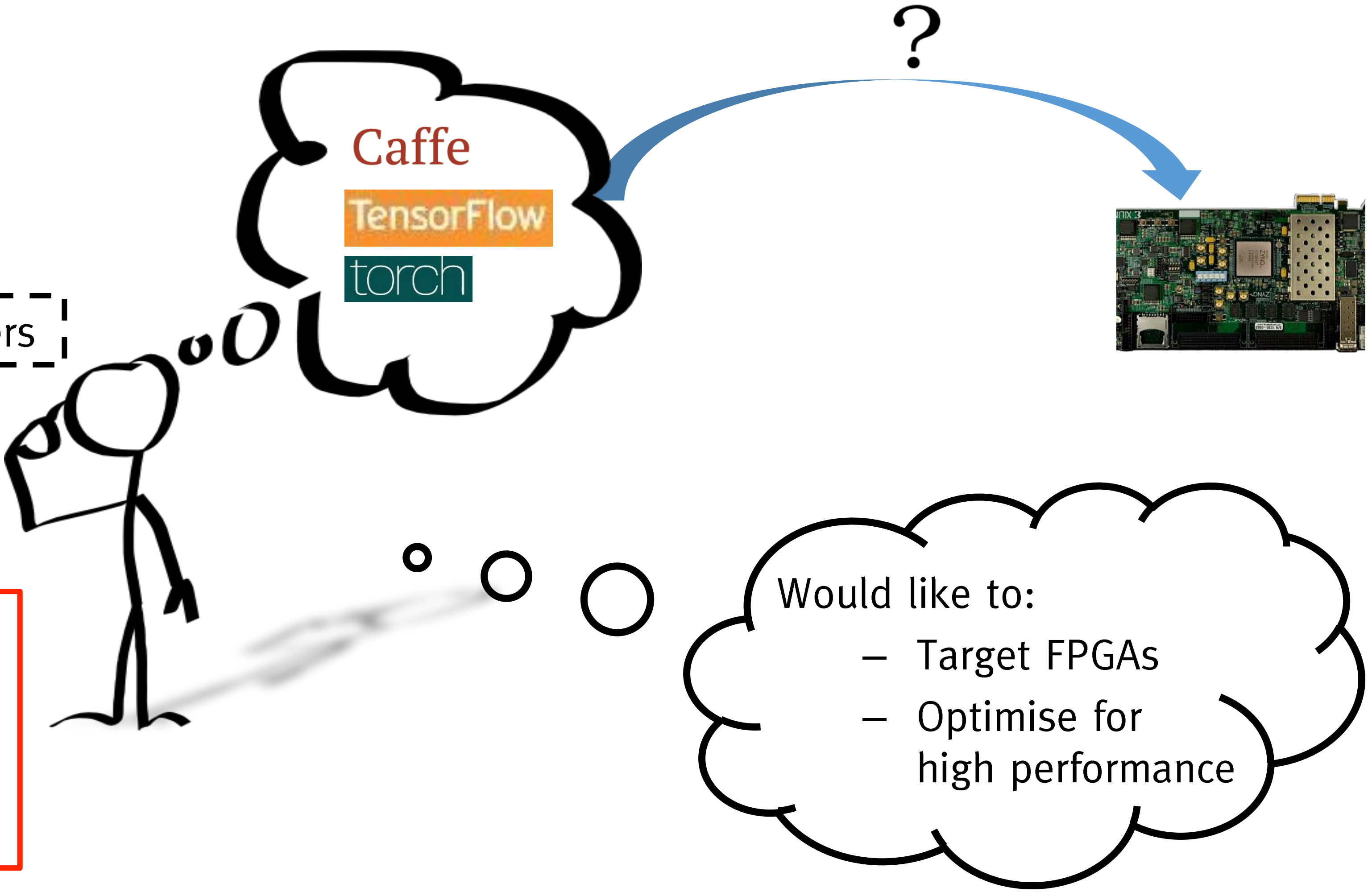
Challenge #1: Mapping Automation



Challenge #1: Mapping Automation

Little knowledge about FPGAs
Ease of deployment
“Good” designs

Deep Learning Developers



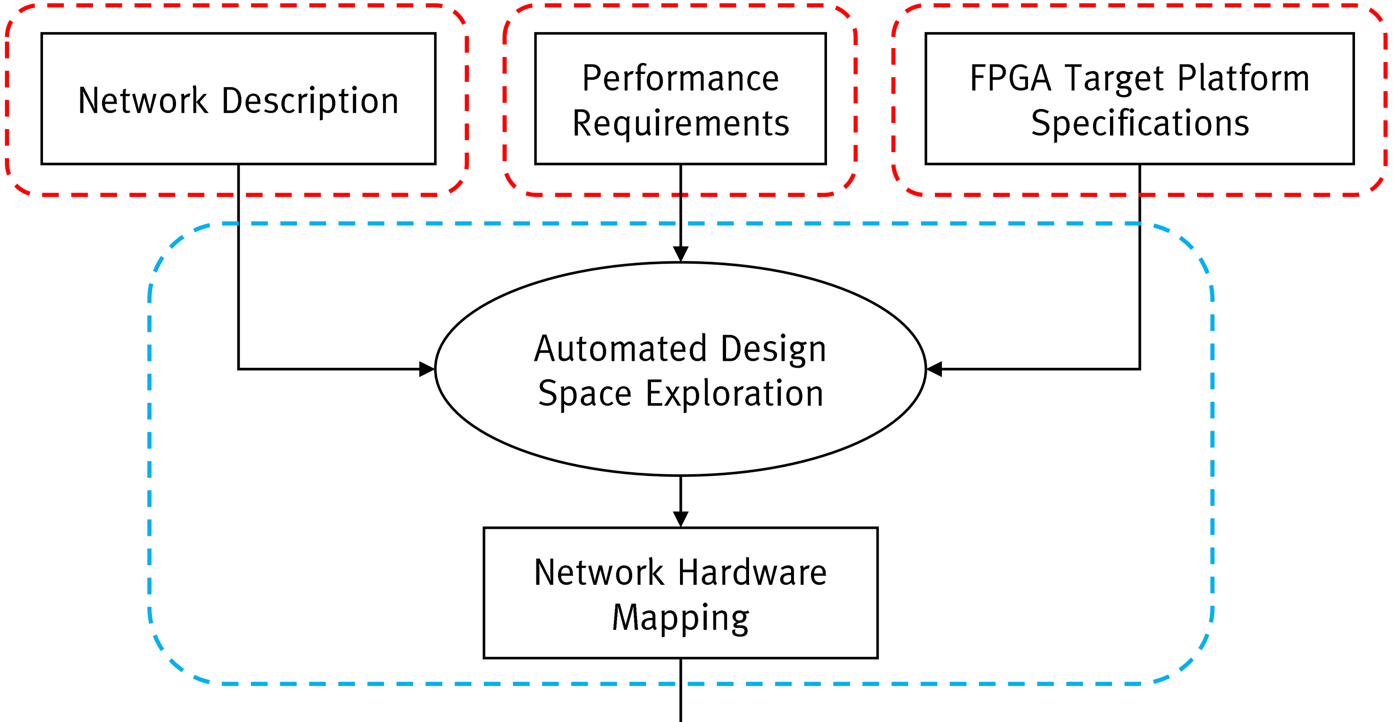
Would like to:
- Target FPGAs
- Optimise for high performance

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

Challenge #1: Automated CNN-to-FPGA Toolflow

Caffe

torch



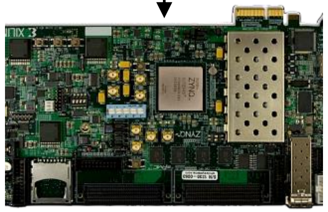
Network Description

Performance Requirements

FPGA Target Platform Specifications

Automated Design Space Exploration

Network Hardware Mapping



Supplied by Deep Learning Expert

fpgaConvNet

fpgaConvNet – Design Space Exploration and Optimisation

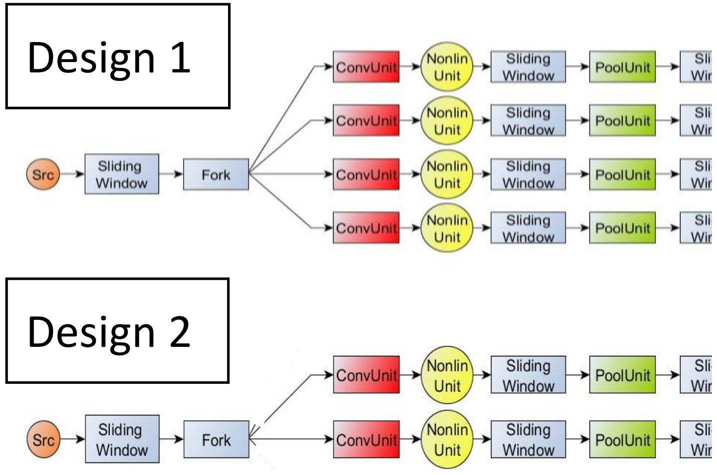
- Synchronous Dataflow Modelling
  - Capture hardware mappings as matrices
  - Transformations as *algebraic operations*
  - Analytical *performance model*
  - Cast design space exploration as a mathematical optimisation problem

Hardware Stages

Interconnections

Design 1

Design 2

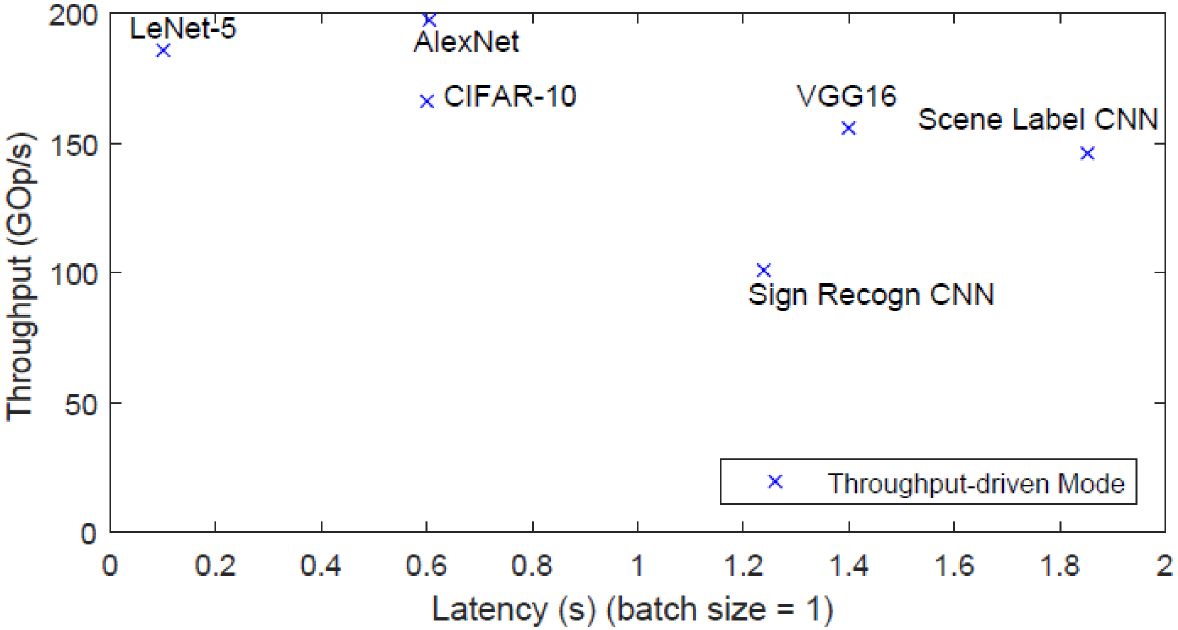


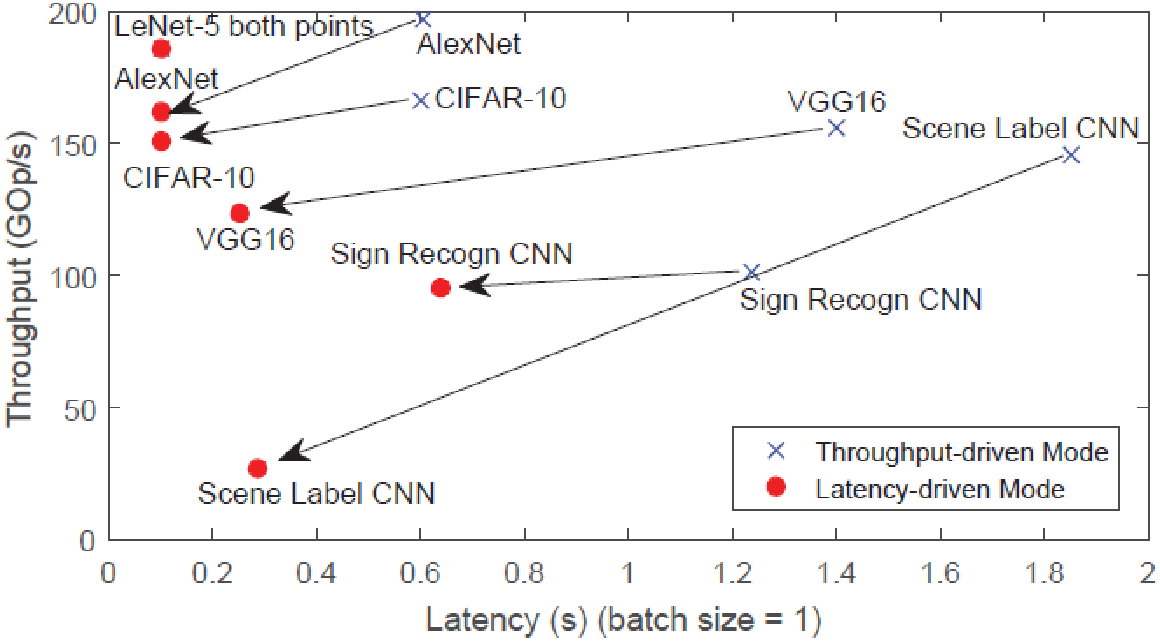
$$\Gamma_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K^2 & -K^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4K^2 & -4K^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4P^2 & -4P^2 \end{bmatrix}$$

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

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

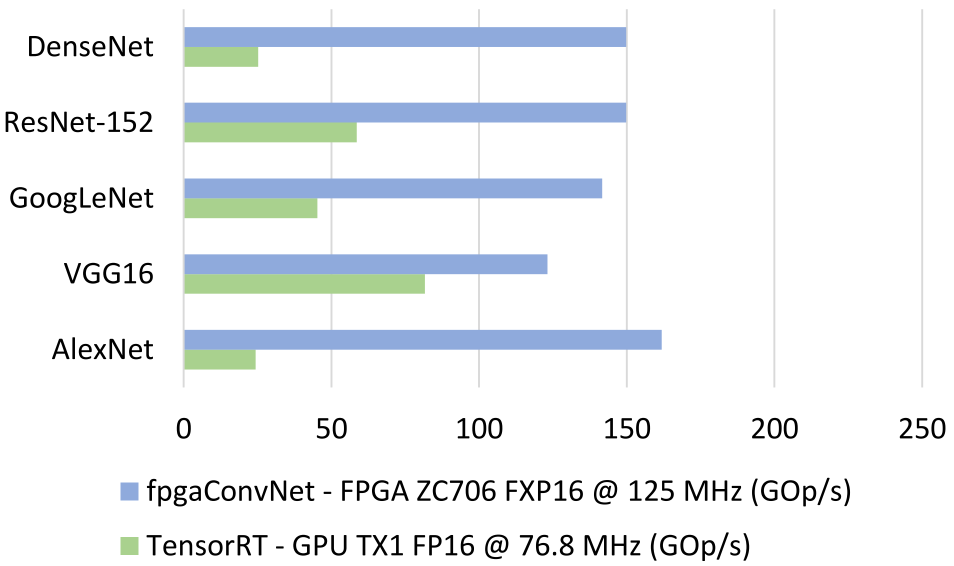
Meeting the performance requirements

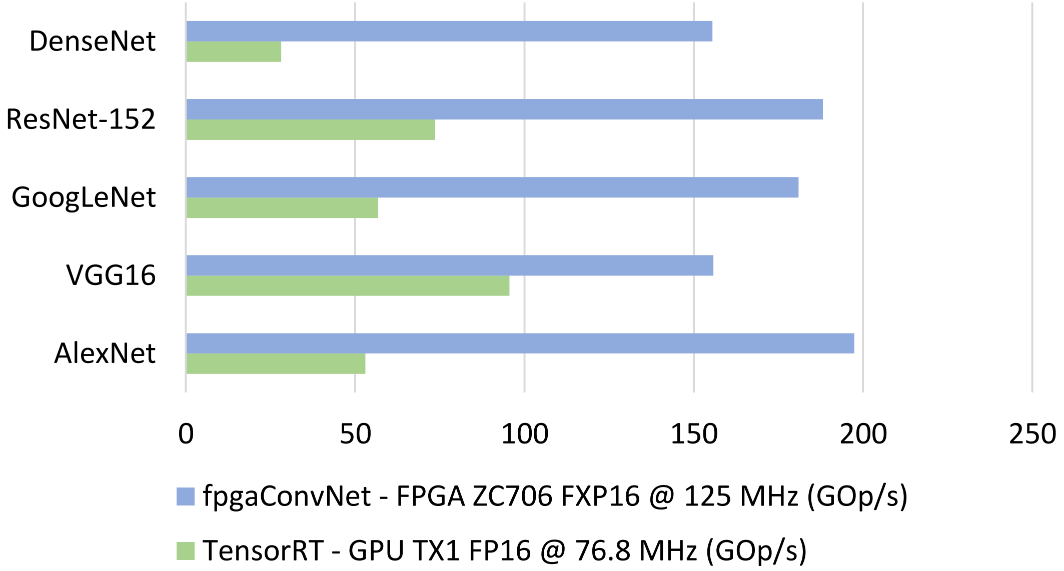




Comparison with Embedded GPUs: Same absolute power constraints (5W)

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



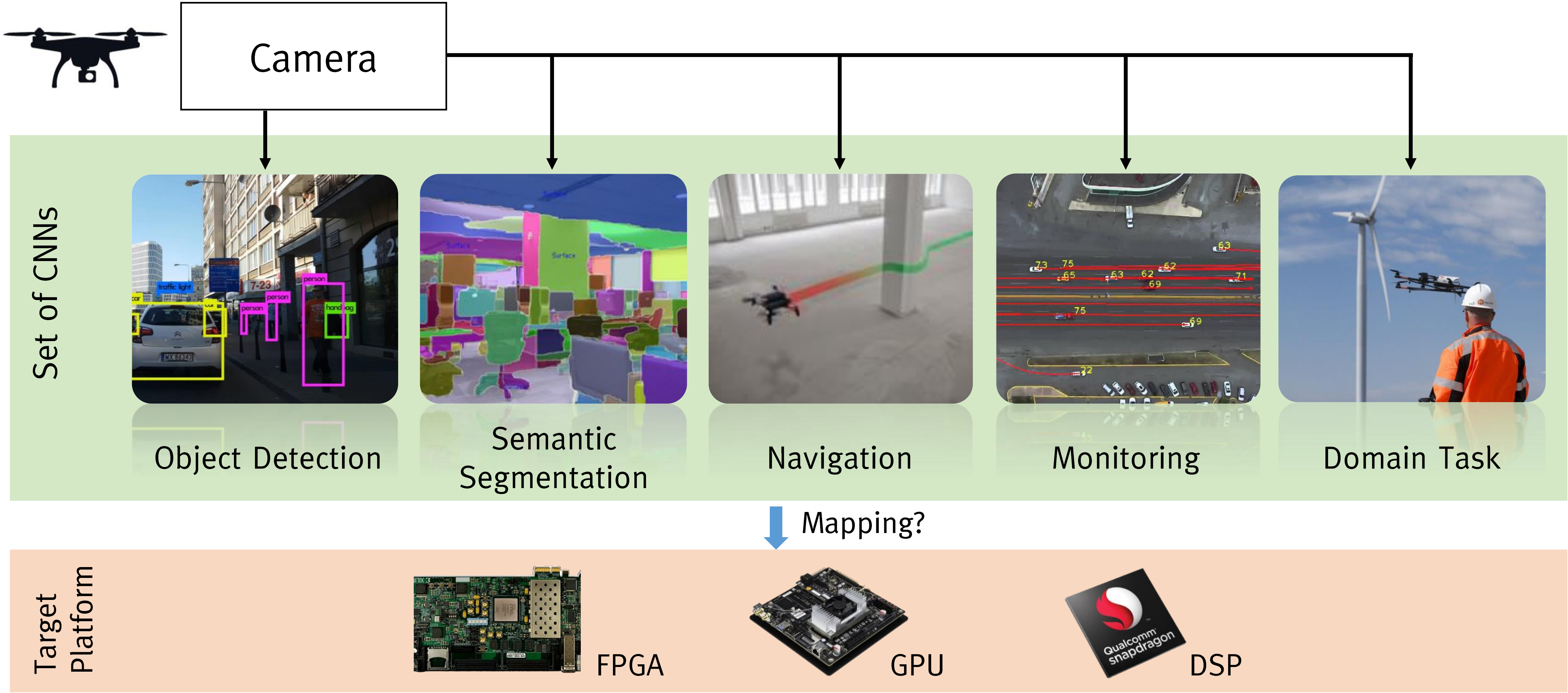


- Latency-driven scenario → batch size of 1
- Up to 6.65× speedup with an average of 3.95× (3.43× geo. mean)

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

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

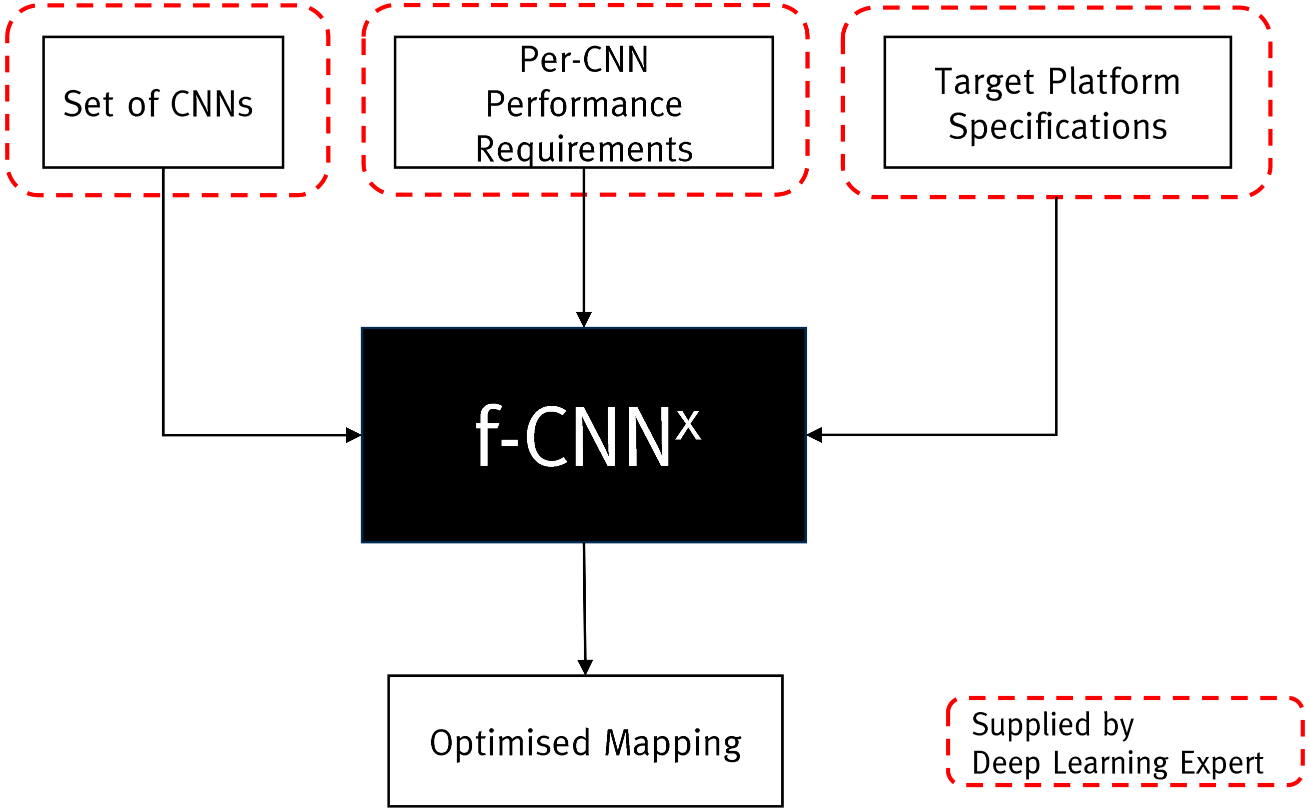
Challenge #2: Multi-CNN System

Challenges:

- Resource allocation among CNNs
- Design automation

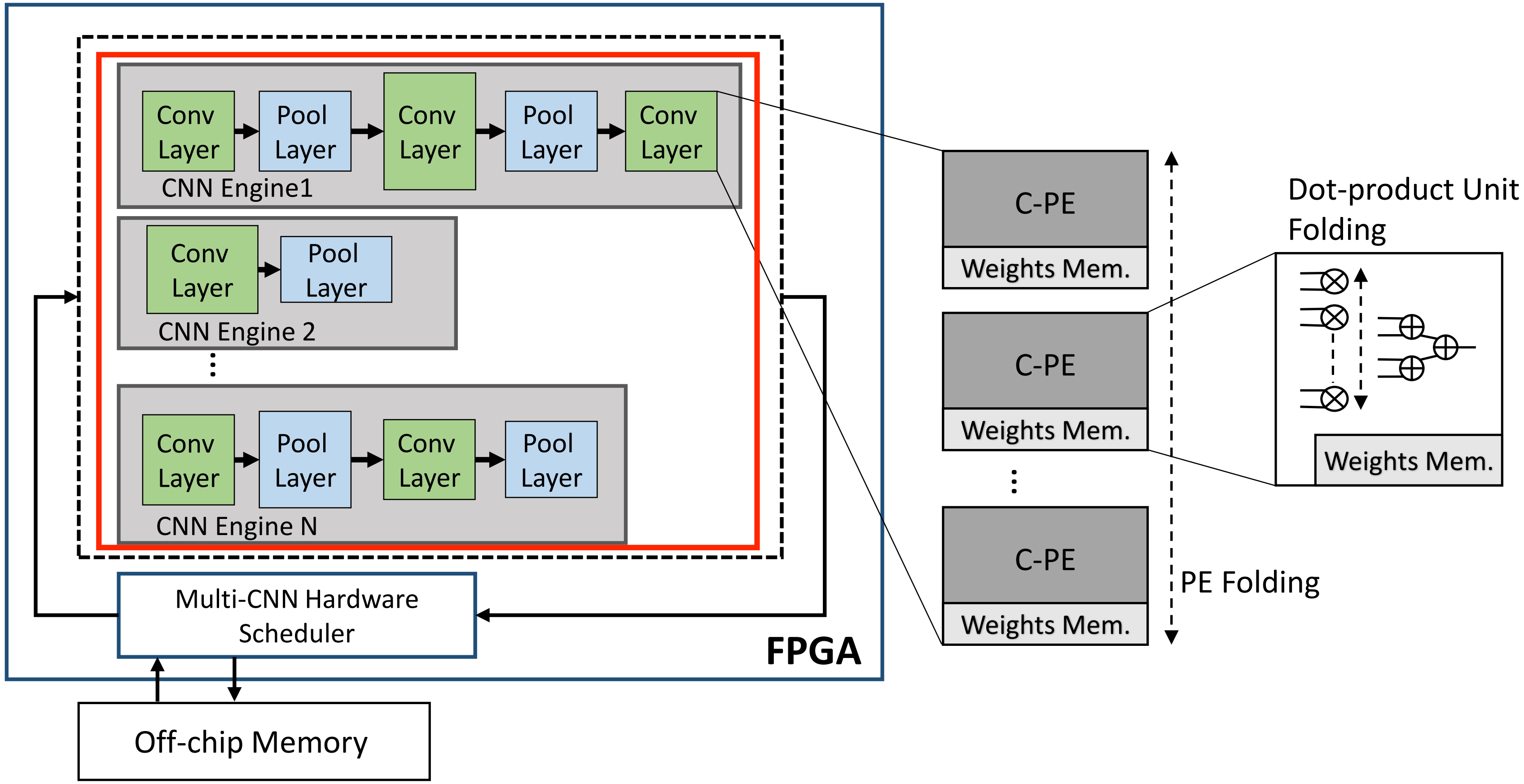
Why?

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

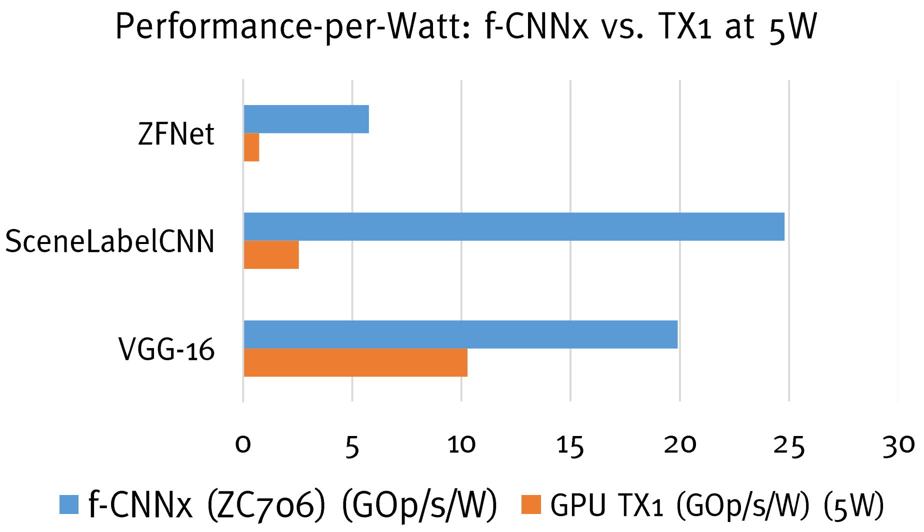

Set of CNNs
Per-CNN Performance Requirements
Target Platform Specifications
f-CNN^x
Optimised Mapping
Supplied by Deep Learning Expert

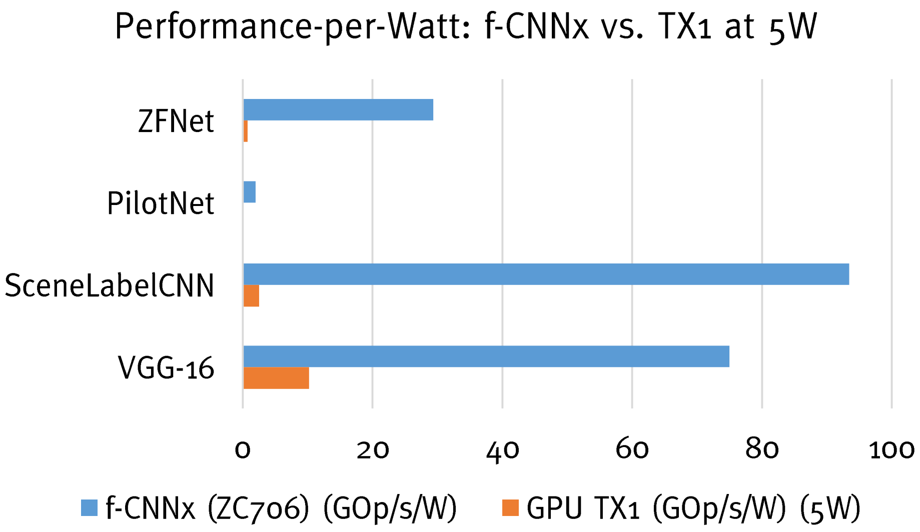
Multi-CNN FPGA design

- One customised hardware engine per CNN
- Explore both on-chip resource allocation and different memory access schedules



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



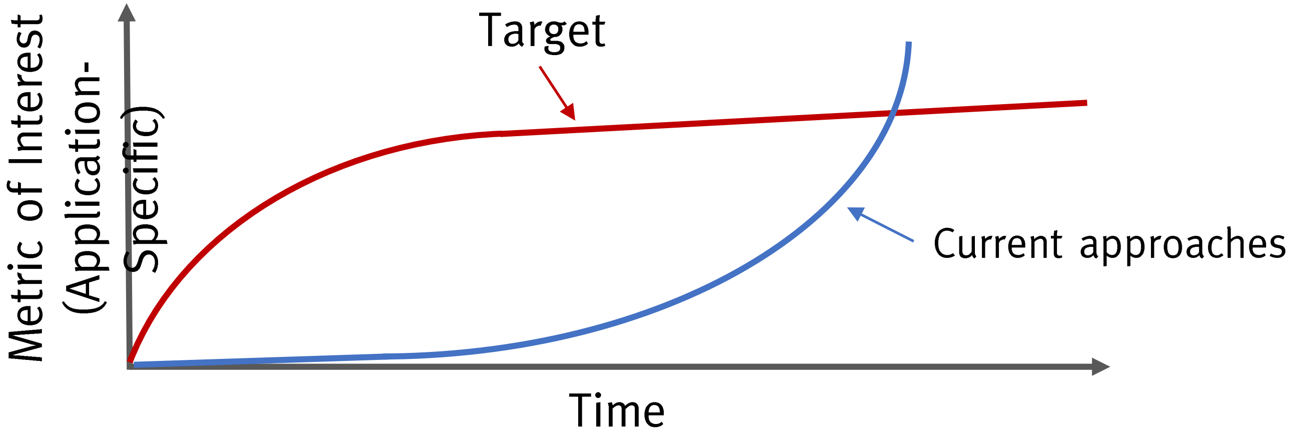


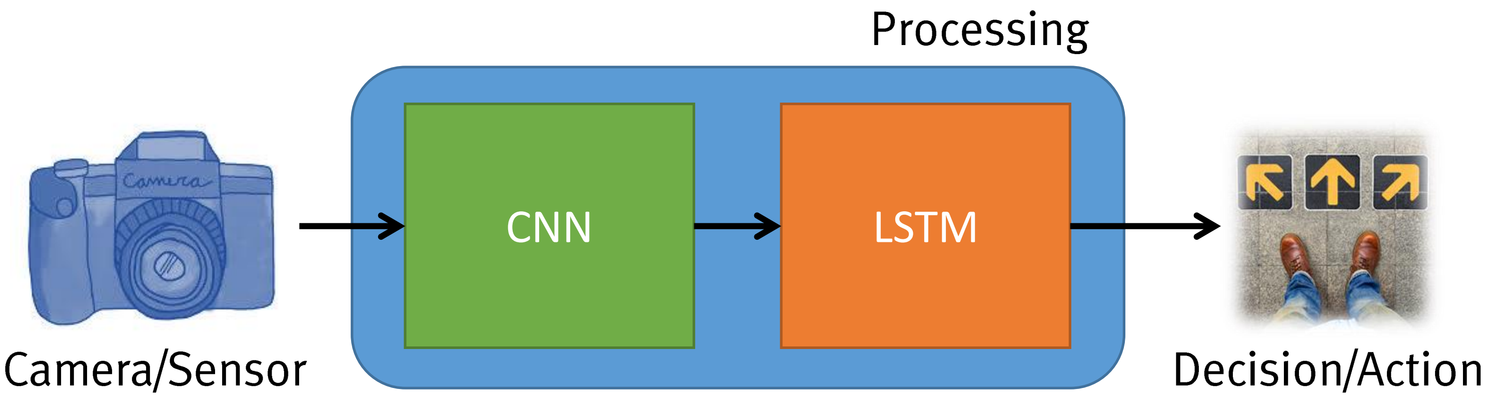
- Latency-driven scenario → batch size of 1
- Up to 9.68× speedup with an average of 5.25× (geo. mean)

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

Challenge #3: Time-constrained Inference

Challenge #3: Time-constrained Inference



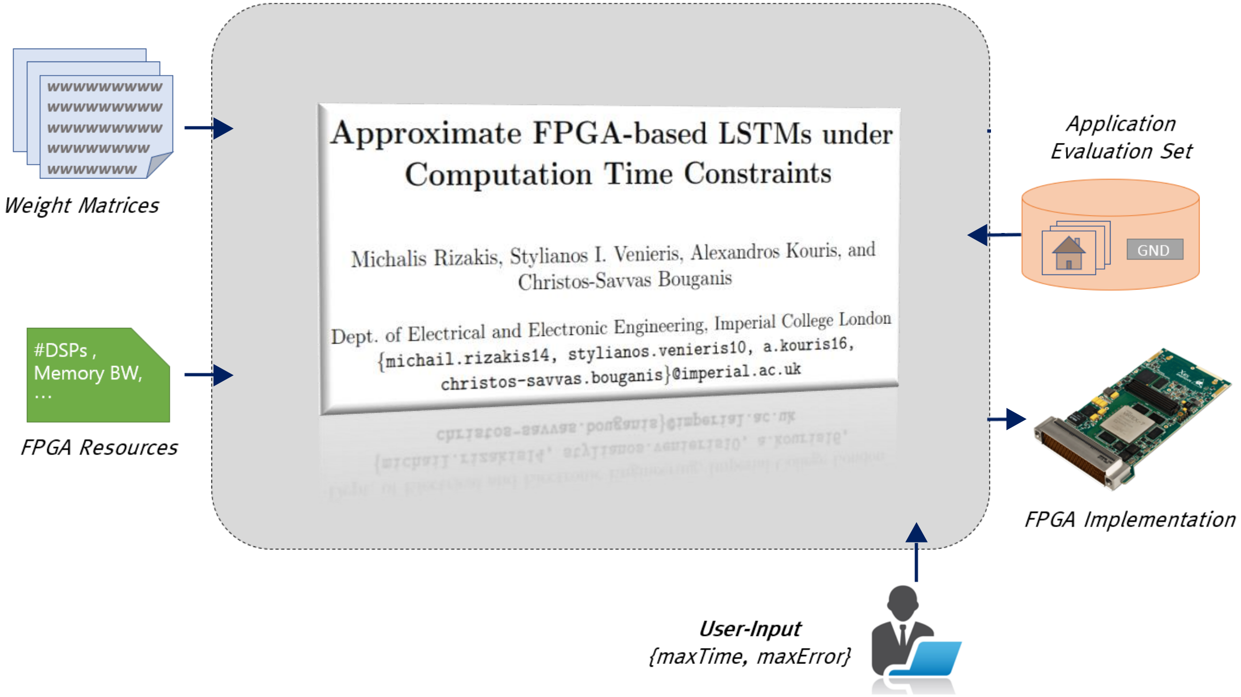


Challenge #3: Time-constrained Inference

- Approximate LSTMs
  - Iterative refinement using SVD + Pruning.
  - Paremetrised with respect to:
    - Number of iterations
    - Level of pruning

- Parametrised hardware architecture, tailored for approximate LSTMs

- Co-optimise given a user-defined time budget



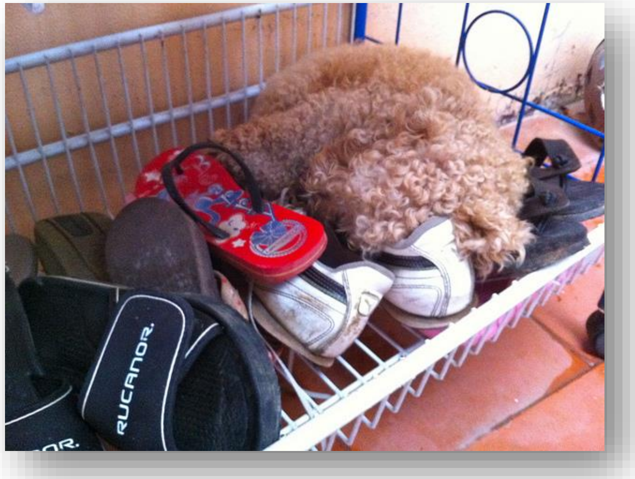
Approximate FPGA-based LSTMs under Computation Time Constraints

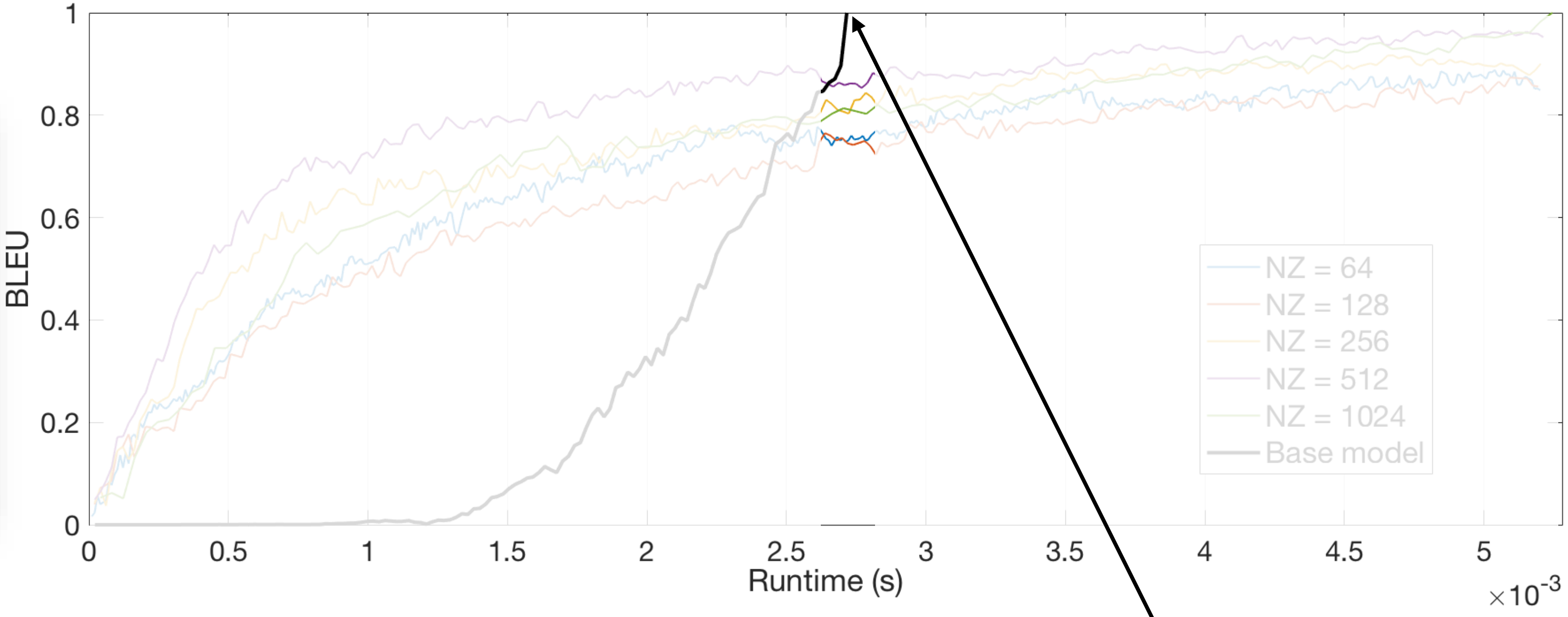
Michalis Rizakis, Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis

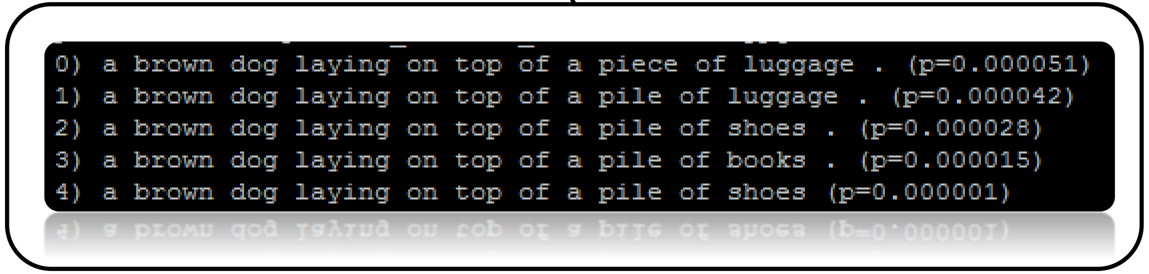
Dept. of Electrical and Electronic Engineering, Imperial College London
{michail.rizakis14, stylianos.venieris10, a.kouris16, christos-savvas.bouganis}@imperial.ac.uk

Impact on LSTM-based Image Captioning

Input Image



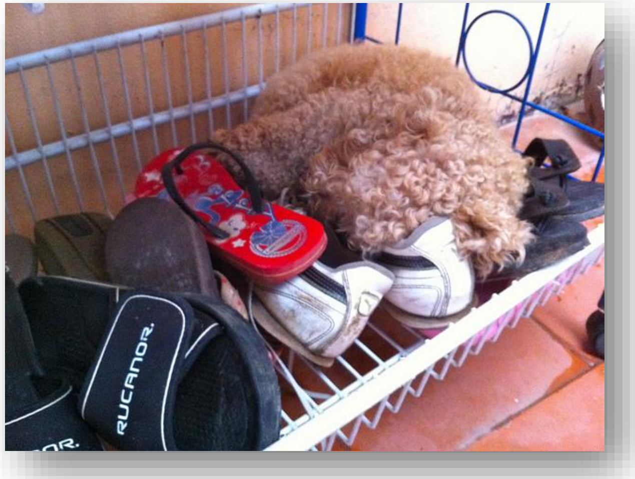


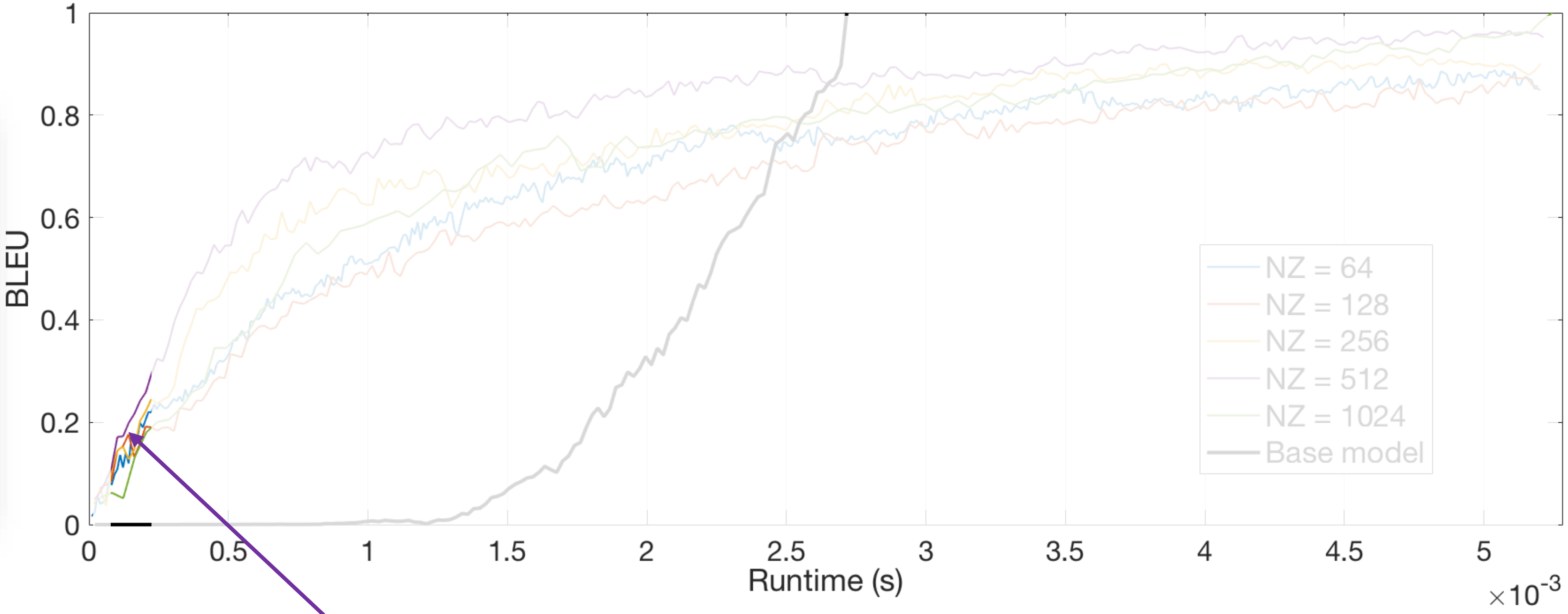


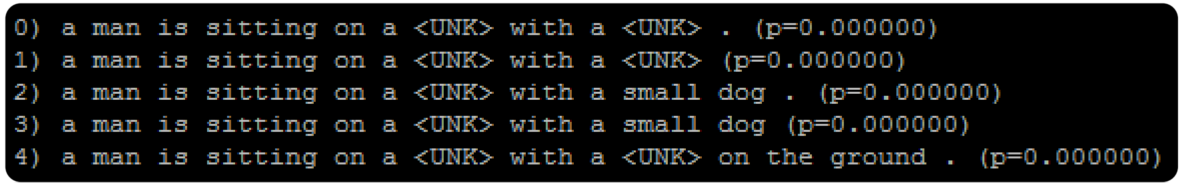
```
0) a brown dog laying on top of a piece of luggage . (p=0.000051)
1) a brown dog laying on top of a pile of luggage . (p=0.000042)
2) a brown dog laying on top of a pile of shoes . (p=0.000028)
3) a brown dog laying on top of a pile of books . (p=0.000015)
4) a brown dog laying on top of a pile of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

Impact on LSTM-based Image Captioning

Input Image



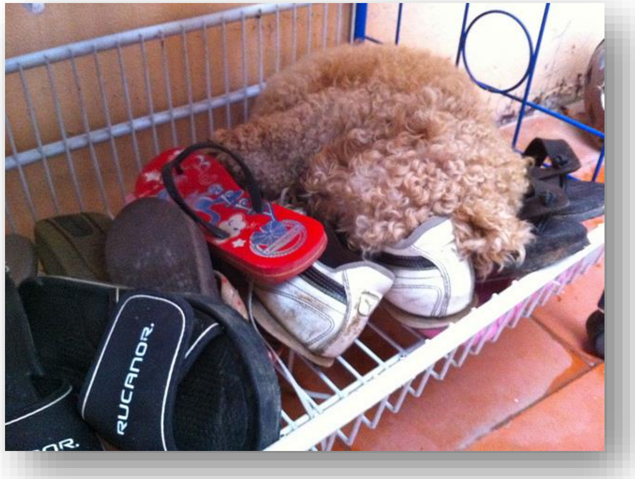


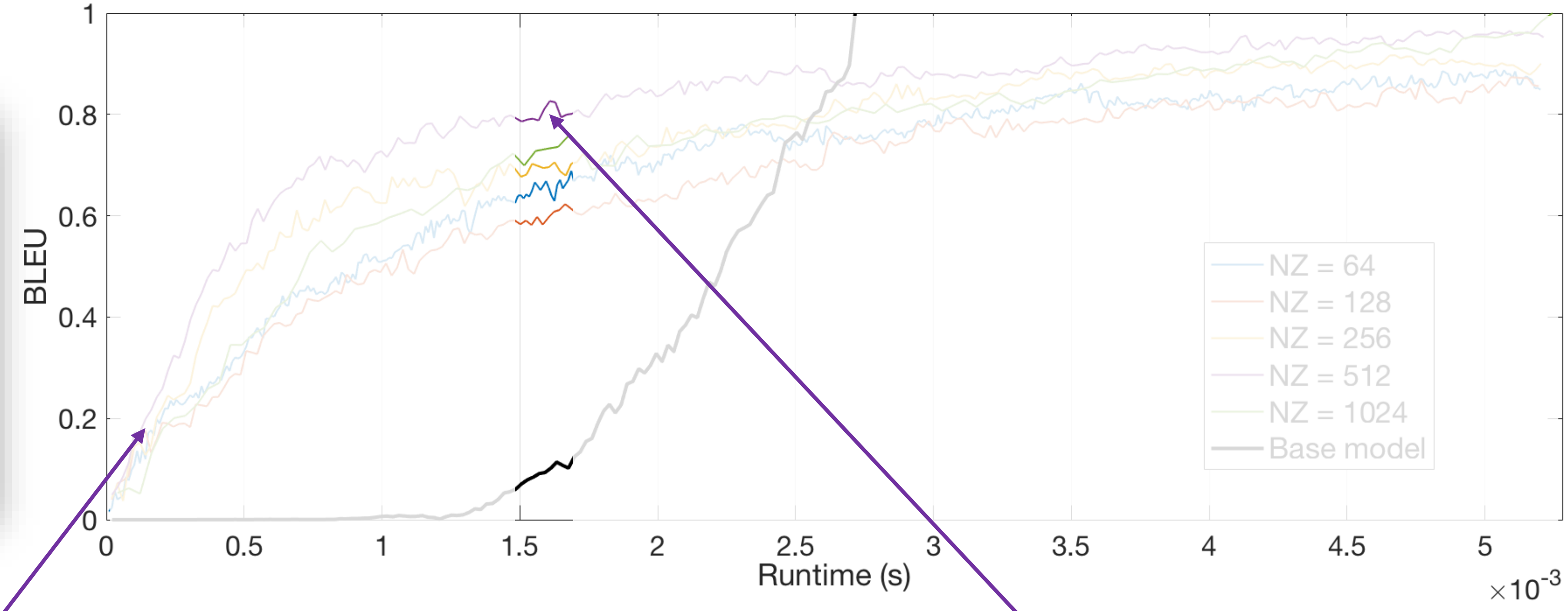


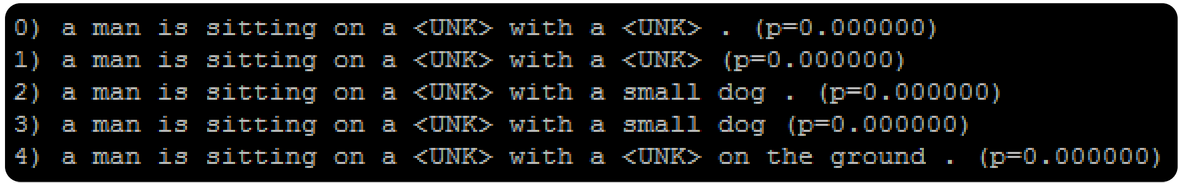
```
0) a man is sitting on a <UNK> with a <UNK> . (p=0.000000)
1) a man is sitting on a <UNK> with a <UNK> (p=0.000000)
2) a man is sitting on a <UNK> with a small dog . (p=0.000000)
3) a man is sitting on a <UNK> with a small dog (p=0.000000)
4) a man is sitting on a <UNK> with a <UNK> on the ground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

Impact on LSTM-based Image Captioning

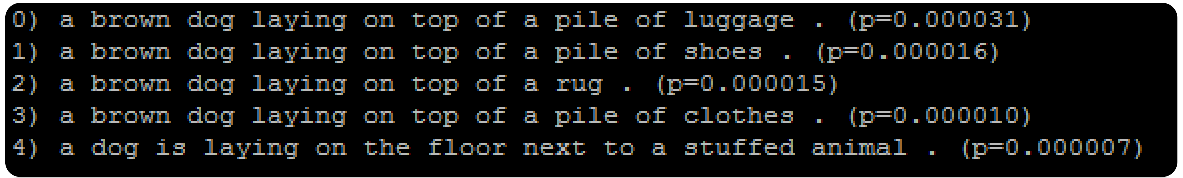
Input Image







```
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

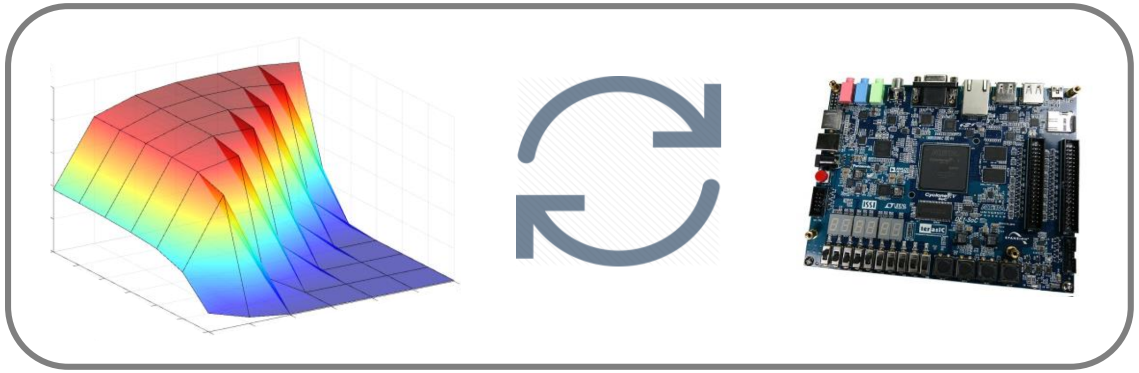
Aim: Design an optimised HW system (performance and accuracy)

Given:

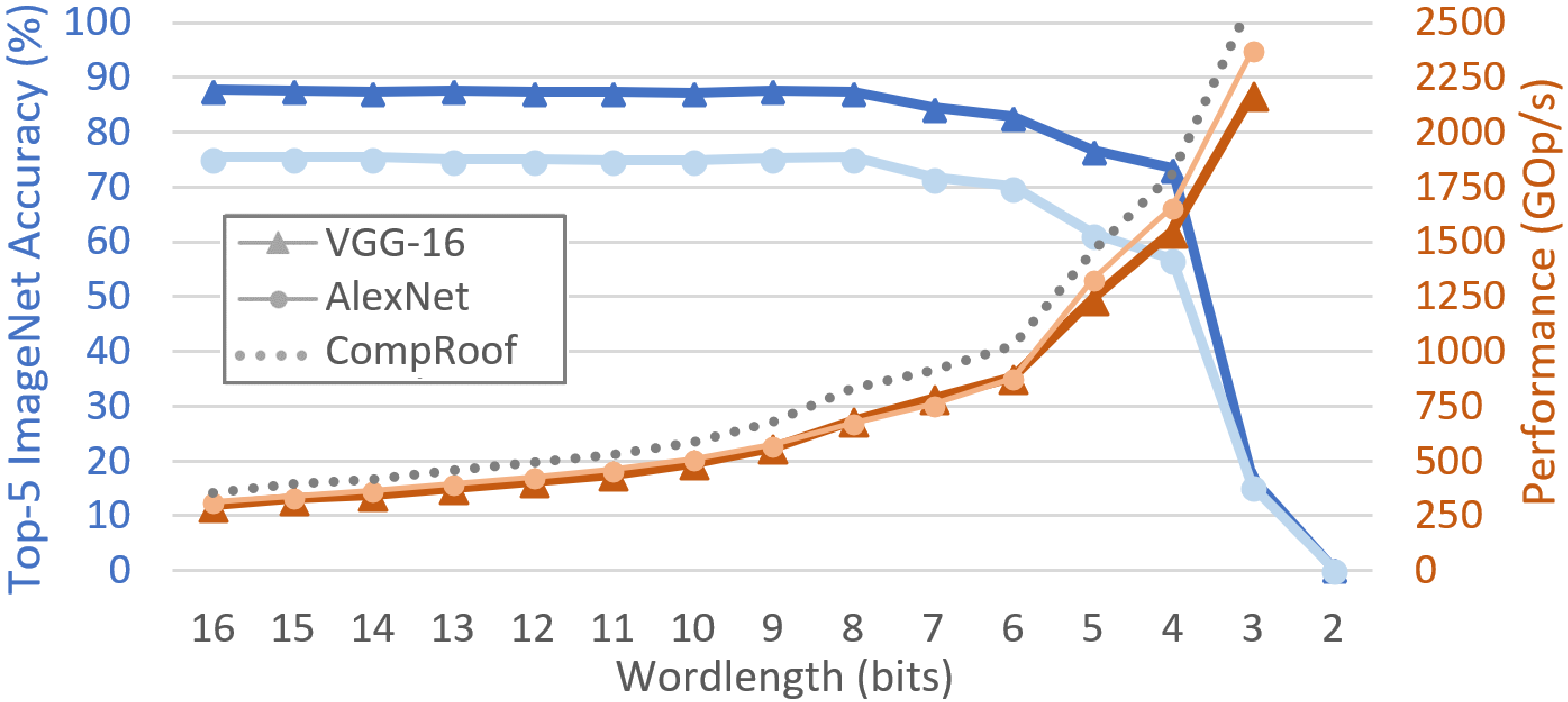
- A High-Level CNN Description (i.e. Caffe)
- A target FPGA platform
- *Train~~ing~~ Data* *privacy, availability*
- *Testing Data*
- Target metric (top1/top-5 accuracy, …)

➔ *quantisation with retraining step*

Limited quantisation opportunities

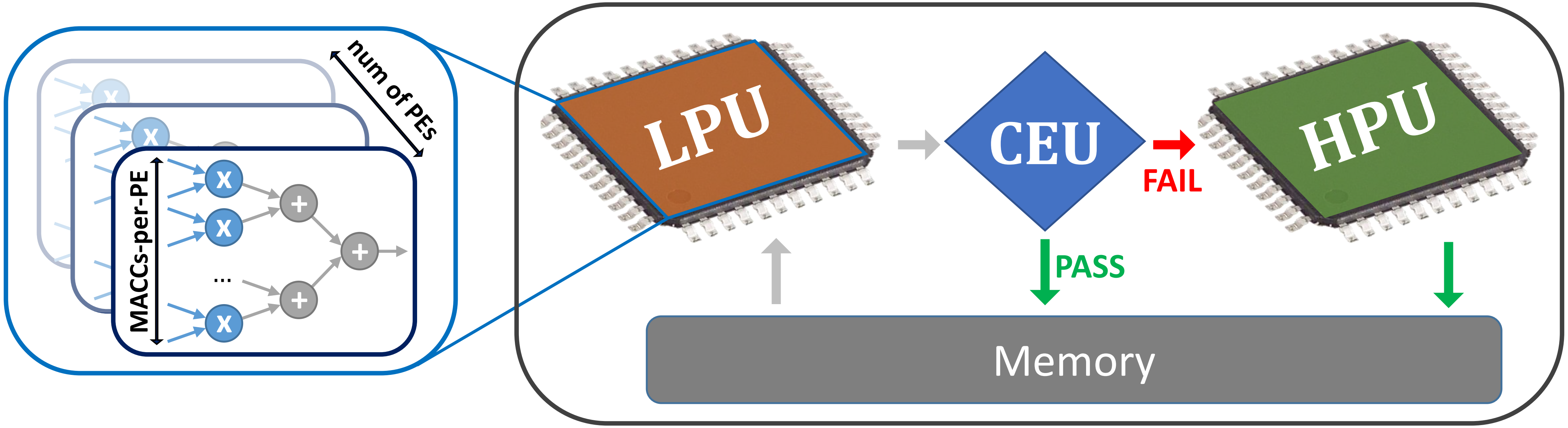


Challenge #4: Privacy-aware Deep Learning



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

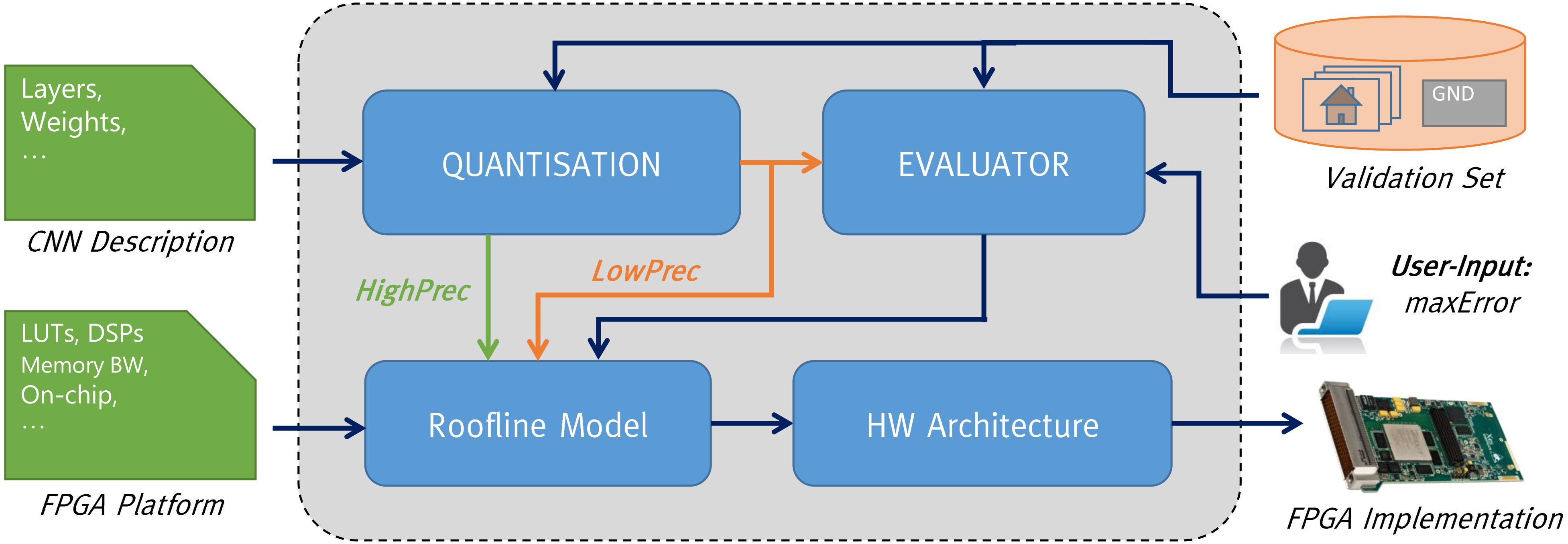


Low-Precision Unit: Degraded accuracy classification with high performance

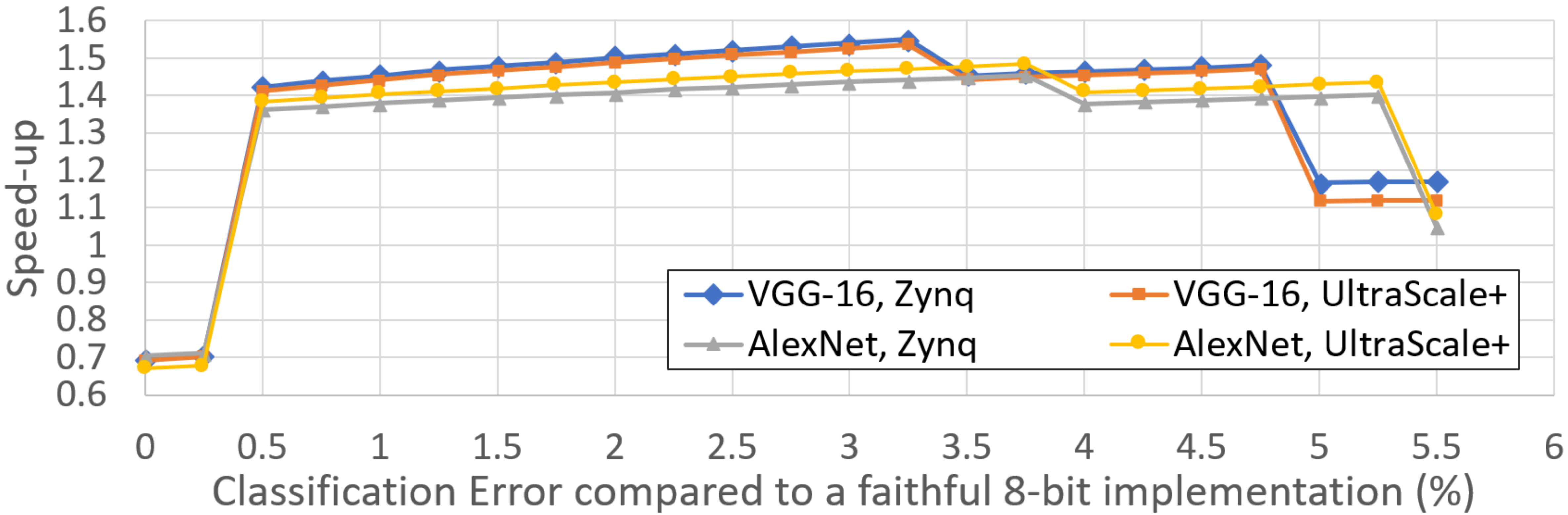
Confidence Evaluation Unit: Identify misclassified cases

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

Challenge #4: Privacy-aware Deep Learning



Challenge #4: Privacy-aware Deep Learning



Summary

Research topics

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

www.imperial.ac.uk/idsl

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.*