Efficient Deployment of Early-Exit DNN Architectures on FPGA Platforms

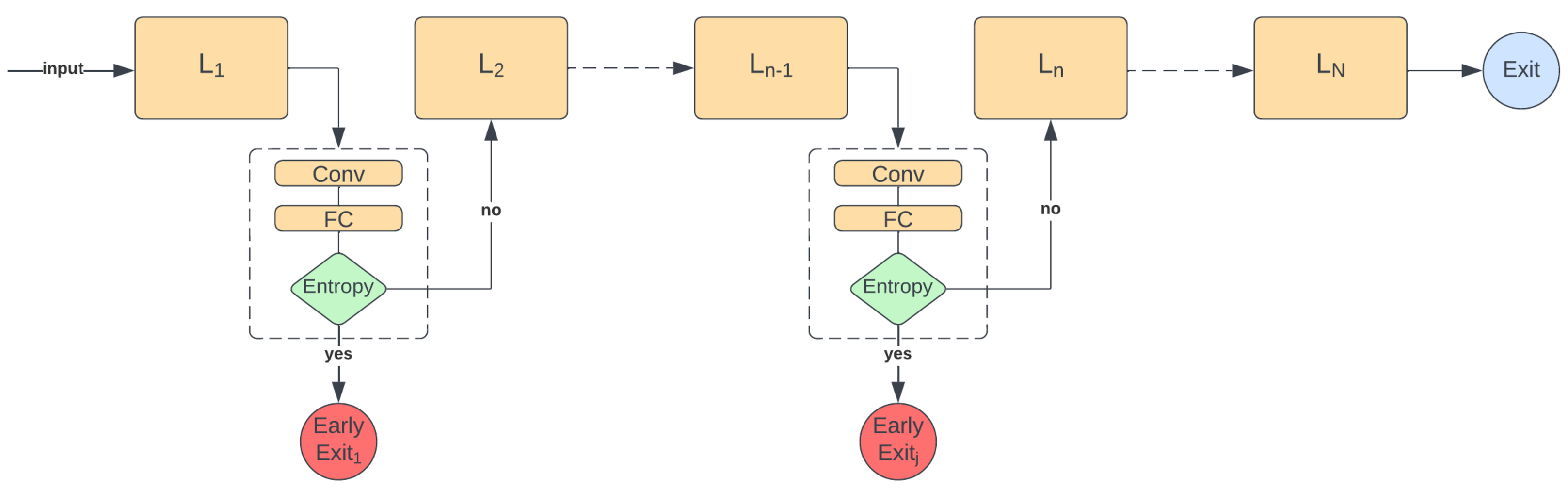
Anastasios Dimitriou, Lei Xun, Jonathon Hare, Geoff Merrett

University of Southampton

Vision of this Thesis

Efficient realisation and implementation of Early-Exit DNNs on FPGAs

- Eliminate unnecessary computations, without sacrificing data.
- Decrease average energy consumption and execution time.
- Enable the deployment of DNNs on resource restricted devices.



Pipeline approach for the design of the decision sub-networks of dynamic DNNs.

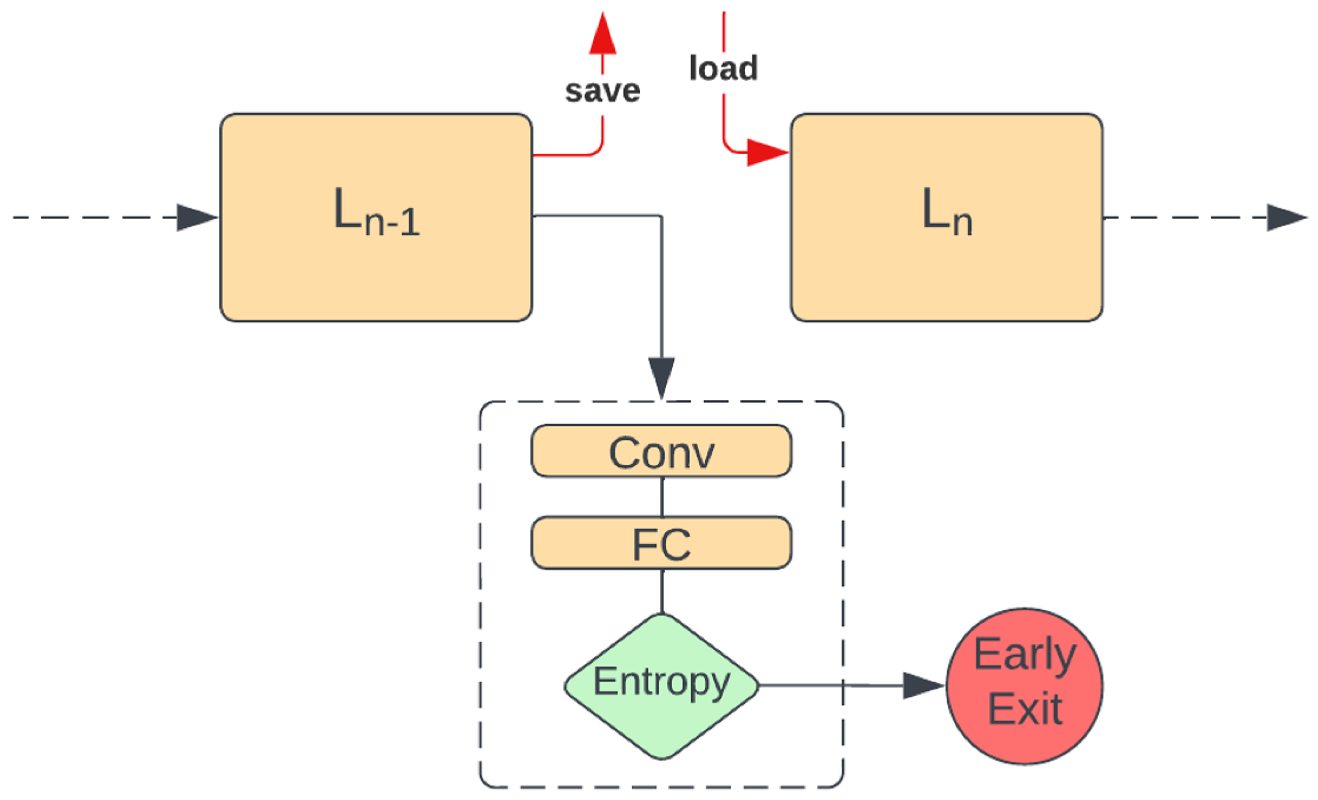
Limitation in Designing

- Mapping tool-flows produce inaccessible designs that are integrated with static network libraries → design from scratch using VHDL.
- Early-Exit DNNs require custom layer types and don't have standard data-paths → custom data loading patterns.
- Halting execution increases use of buffering and external memory → parallel approach.

Proposed Architectures

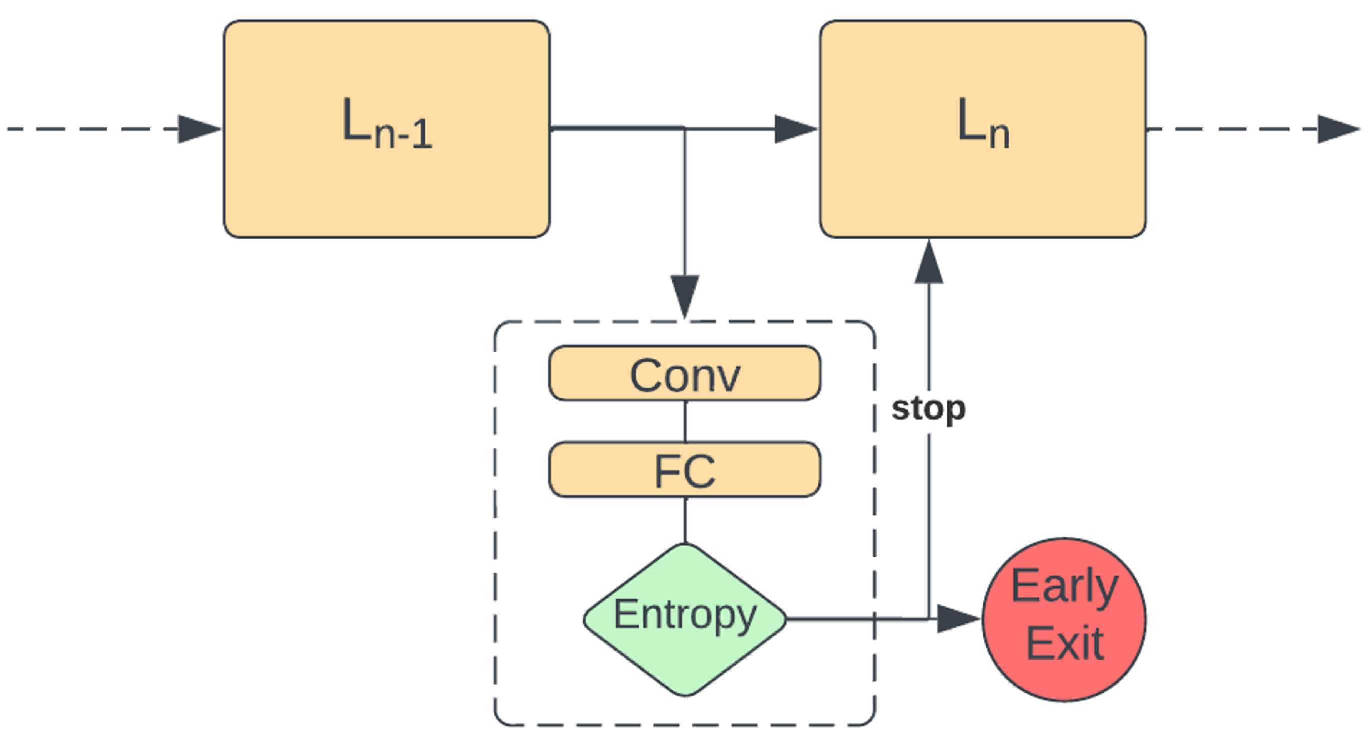
Pipeline Approach

- Opportunity for reusing existing IPs
- Simpler design → lower energy needs
- Requires the intermediate output to be stored in memory
- Increase latency when full depth is needed



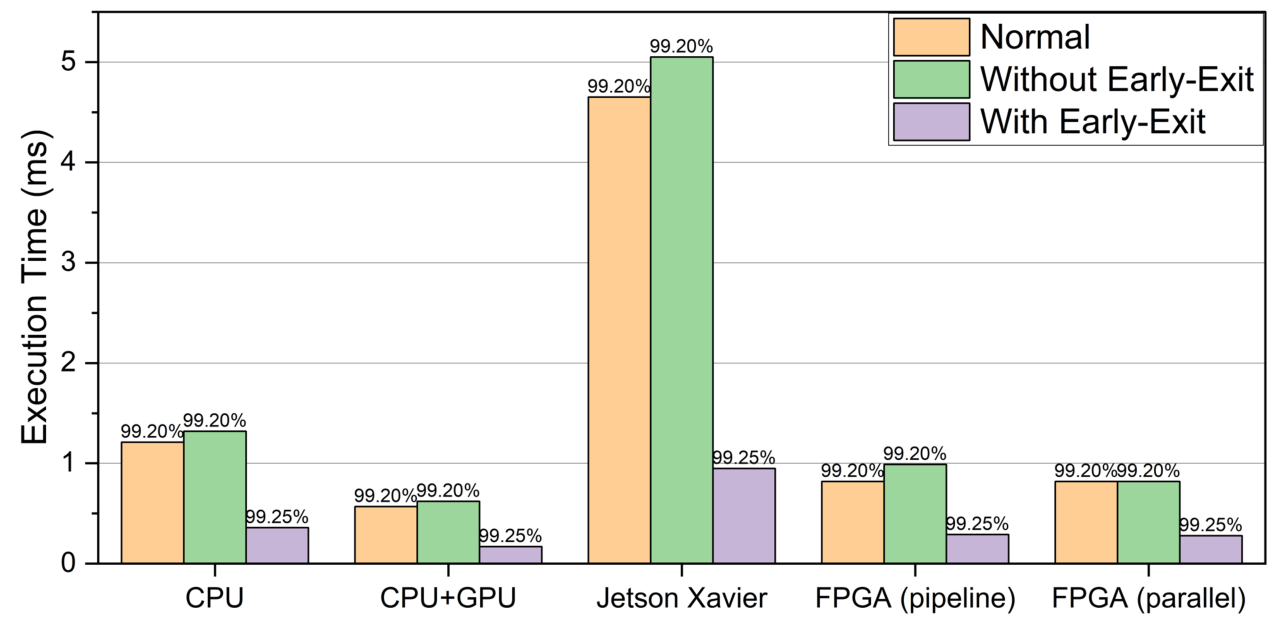
Parallel Approach

- Separate design for the decision sub-network → higher utilization
- No latency drop on the backbone execution
- Lower memory needs
- Complex design → higher energy needs



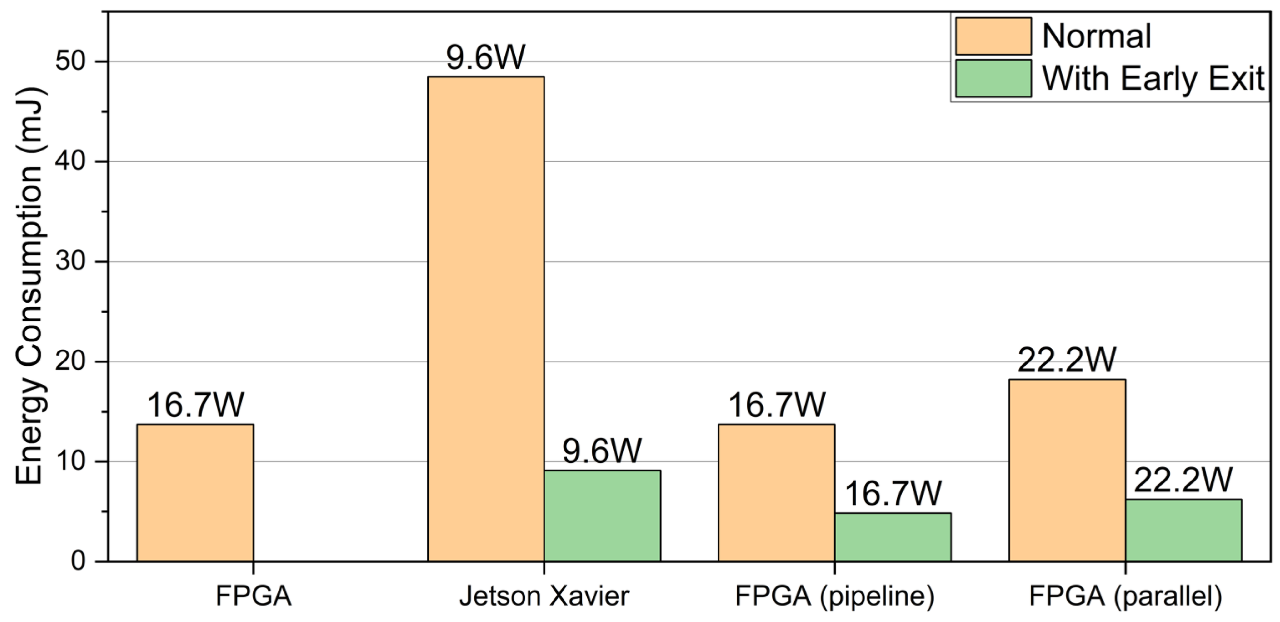
FPGA Accelerator Experimental Results

For the evaluation BranchyNet [1] early exit LeNet-5 DNN is used.



Experimental results comparing the per sample execution time.

- Early-Exiting approach is at least 2.8x faster.
- FPGAs performance is competitive over CPU/GPU.
- *Parallel* approach is **1.2x** faster than the *pipeline* on "hard" samples.

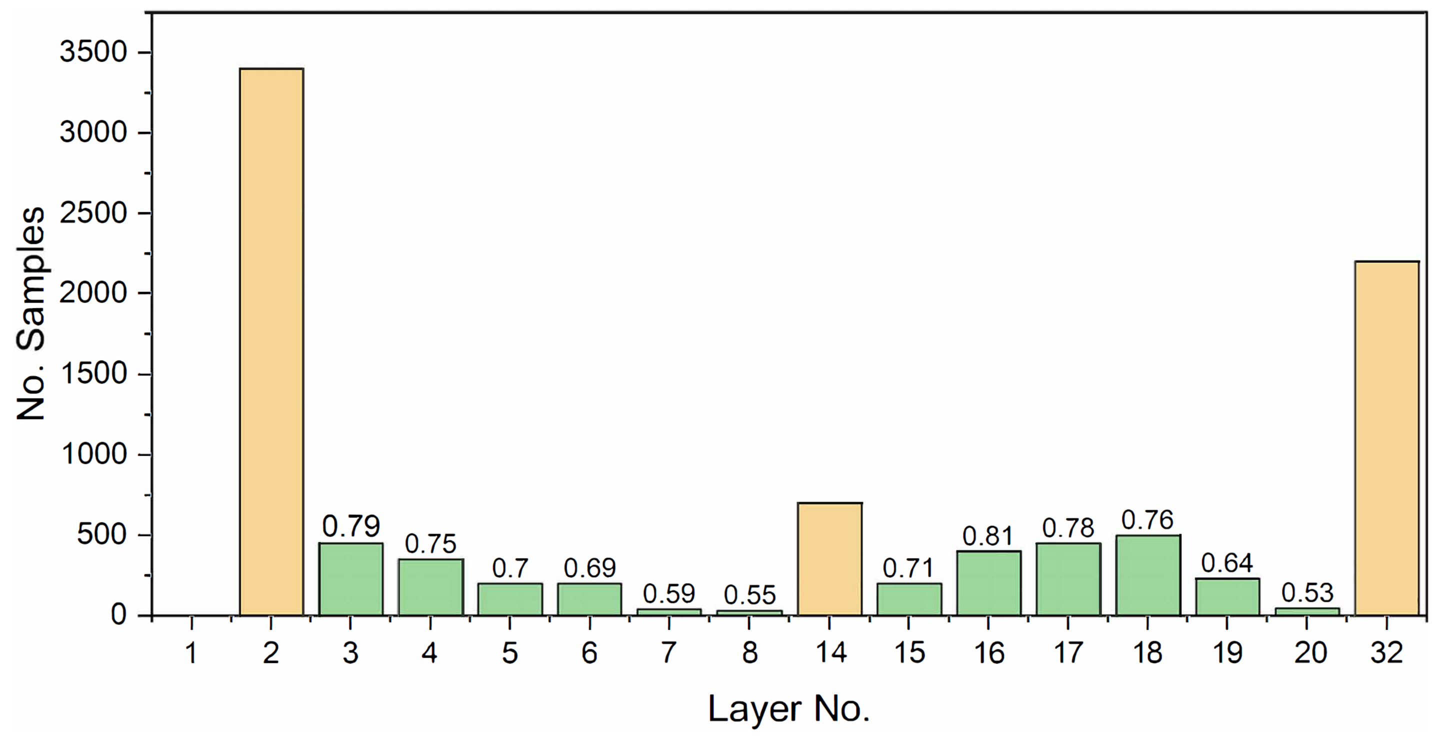


Experimental results comparing the per sample energy consumption.

- Early-Exiting approach is at least 2.8x less energy demanding.
- FPGAs capabilities are highlighted, being at least **1.5x** less energy demanding over Jetson.

Early-Exit Placement Optimization

- Early-exit architecture mandates predetermined and fixed exit points.



Number of samples to be firstly correctly predicted in the next 6 layers after being rejected from the predetermined exit points on ResNet-32.

- **31%** of rejected outputs from **original exit points** recognized by subsequent **6 layers**.
- *Parallel* approach allows dynamic confidence-controlled placement of early-exit points.
- Early-Exit ResNet32 computations reduced by **24%** while maintaining same accuracy levels.

Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under EP/S030069/1. Data associated with this paper is available on https://doi.org/10.5258/SOTON/D2885.

References

[1] S. Teerapittayanon, B. McDanel, H. T. Kung. "BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks.", in Conf. ICPR, 2016.

International Centre for Spatial Computational Learning

Engineering and Physical Sciences Research Council

Author contact details
Email: ad1r20@soton.ac.uk

DATE '24, 25/03/2024, Valencia