fpgaConvNet: Automated Mapping of CNNs on FPGAs
Stylianos I. Venieris and Christos-Savvas Bouganis
Department of Electrical and Electronic Engineering, Imperial College London
{stylianos.venieris10, christos-savvas-bouganis}@imperial.ac.uk

Introduction
- Convolutional Neural Networks (ConvNets) are state-of-the-art in AI tasks, from object recognition to natural language processing.
- Several frameworks have been released which enable faster experimentation and development of ConvNet models by targeting powerful GPUs.
- FPGAs provide a potential alternative in the form of a low-power, reconfigurable platform with tunable trade-offs between throughput, latency, and resource cost.
- fpgaConvNet is an end-to-end tool that aims to bridge the gap between existing Deep Learning toolchains and FPGAs targeting both high-throughput and low-latency applications [1].

Design Space Exploration and Optimisation
- The design space exploration method uses performance and resource models to traverse the design points described by the tunable parameters of the building blocks.
- A set of four transformations over the SDF model are defined:
  - Graph Partitioning with Reconfiguration
  - Coarse-Grained Folding
  - Fine-Grained Folding
  - Weights Reloading
- The optimiser operates in two modes by selecting objective function based on the performance metric of interest: throughput or latency. We pose two combinatorial constrained optimisation problems for high-throughput and low-latency applications respectively:

\[
\min T(B, \Gamma), \text{s.t. } rsc(B, \Gamma) \leq rsc_{\text{At_max}}.
\]

\[
\min L(1, \Gamma), \text{s.t. } rsc(B, \Gamma) \leq rsc_{\text{At_max}}.
\]

where \( T, L, \) and \( rsc \) return the throughput in GOp/s, the latency in s/input and the resource consumption of the current design point, \( B \) is the batch size and \( rsc_{\text{At_max}} \) is the resource budget of the target FPGA.

- The optimiser employs all four SDF transformations to traverse the architectural design space and aims to find a design point that approximately optimises the objective function.

Evaluation
- Throughput-driven optimisation employs batch processing for:
  - weights reuse across inputs in the batch,
  - amortisation of the FPGA reconfiguration and reloading of the weights.
- Latency-driven design shifts the generated design points to low-latency regions with an average latency improvement of 25.4x.
- fpgaConvNet achieves 2.5x and 1.5x higher throughput and throughput density compared to the AlexNet design by Zhang et al. [5]. Moreover, our framework achieves 1.25x and 3.98x higher throughput and throughput density compared to the OpenCL-based VGG16 design by Suda et al. [6].

References