Project Proposal Outline

RESEARCH INTERNSHIP

Indian Institute of Science

Research Guide-

Prof. L. M. Patnaik,

Adjunct Professor and INSA Senior Scientist,

NIAS, Indian Institute of Science.

Hardware Acceleration for Convolutional Neural Networks on FPGA.

Introduction: The accuracy of convolutional neural networks has been continuously improving, but the computational cost of these networks also increases significantly. Recent breakthroughs in the development of multi-layer convolutional neural networks have led to state of the art improvements in the accuracy of non-trivial recognition tasks such as large-category image classification and automatic speech recognition. These many-layered neural networks are large, complex, and require substantial computing resources to train and evaluate. Therefore, choosing a proper computation platform for neural-network-based applications is essential.

Problem Statement: Modern deep learning networks need to perform an enormous amount of floating point computation during both training and inference. A typical CPU can perform 10-100G FLOP per second, and the power efficiency is usually below 1GOP/J. So CPUs are hard to meet the high performance requirements in cloud applications nor the low power requirements in mobile applications. A vast majority of these operations can be computed in parallel, and this is why GPUs are usually the target devices for these tasks. For very deep neural networks, multiple GPUs can be used in parallel. Unfortunately, a single GPU's power consumption can be about 200/300W and its weight about 1kg. These are not good specifications for embedded or mobile devices.

Solution Approach: Field programmable gate arrays (FPGAs) have been extensively studied as an important hardware platform for CNN computations. A lot of research has been done on implementing neural networks on power-friendly devices like FPGAs. Field Programmable Gate Arrays are small devices with little power consumption (about 1/2W) with

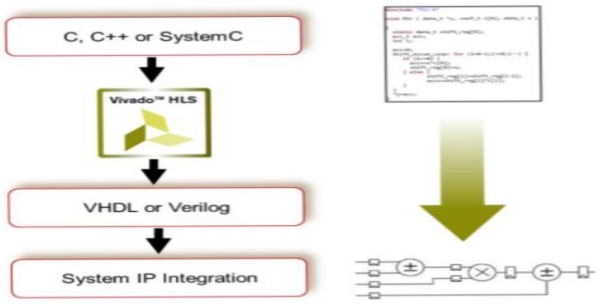
reconfigurable hardware. FPGAs don't have the same speed nor the same amount of resources of a modern GPU, but an efficient implementation could obtain the same throughput (images/sec) of a GPU by using a significantly smaller amount of power. FPGAs are becoming a platform candidate to achieve energy efficient neural network processing. With a neural network oriented hardware design, FPGAs can implement high parallelism and make use of the properties of neural network computation to remove additional logic.

Objective: The long term goal of this Research is to optimise and accelerate the Convolutional Neural Networks and show how FPGAs offer significant efficiency improvements over CPU and GPU. The most important pragmas to be taken care off are:

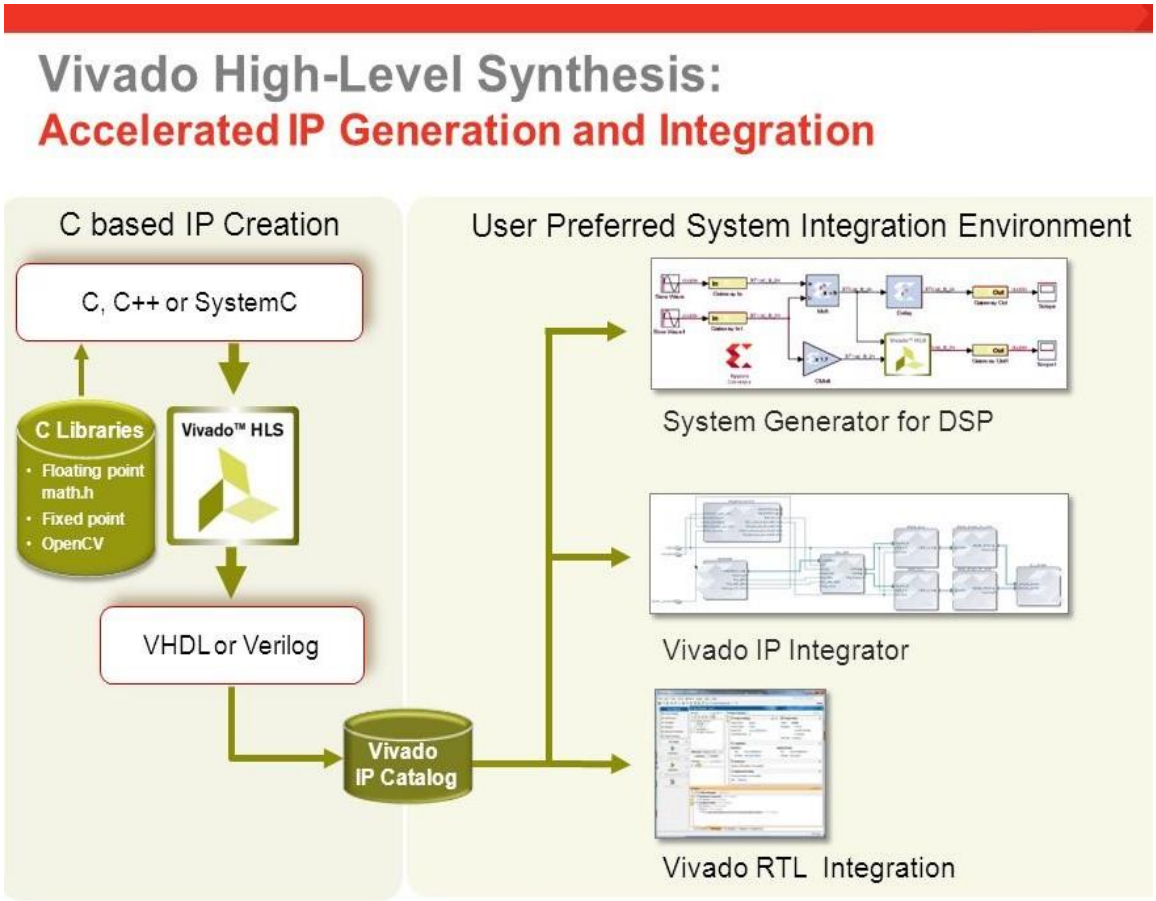
- To infer pipelining into our modules.
- To unroll loops and achieve parallelism.
- To control input and output ports via standard protocols (AXI, AXI-Stream).
- Estimate the Resources and timing of the implemented CNN.

Period of the research: 4th July to 6th September 2019.

Resources required for the research: Virtex-6 FPGA, Xilinx-Vivado HLS, Vivado, SDK tools.



Methodology : For this project we are going to use Xilinx's High-Level Synthesis Tool, Vivado HLS, that allows us to write C++ code that will be translated into RTL code (VHDL, Verilog); the major benefits of using Vivado HLS are its pragams, that allow the programmer to control resource usage, timing requirements and the architecture implementation. we develop a distributed architecture composed of the customized processing elements (PEs) that enables high computation parallelism and data reuse rate of the compressed network.



-Darshan C Ganji,

Electronics and Communication Engineering,

Acharya Institute of Technology.