Deep convolutional neural network has taken an important role in machine learning algorithm which has been widely used in computer vision tasks. However, its enormous model size and massive computation cost have became the main obstacle for deployment of such powerful algorithm in low power and resource limited embedded system, such as FPGA. Recent works have shown the binarized neural networks (BNN), utilizing binarized (i.e. +1 and -1) convolution kernel and binary activation function, can significantly reduce the model size and computation complexity, which paves a new road for energy-efficient FPGA implementation.

This thesis proposes to implement a new parallel binarized convolutional neural network (i.e. PCâ€“BNN) on FPGA with accurate inference. The embedded PCâ€“BNN is designed for image classification on CIFARâ€“10 dataset and explores the hardware architecture and optimization of customized CNN topology.

The PCâ€“BNN replaces the original binary convolution layer in conventional BNN with two parallel binary convolution layers. PCâ€“BNN achieves 86% on CIFARâ€“10 dataset with only 2.3Mb parameter size. We deploy our proposed PCâ€“BNN into the Xilinx PYNQ Z1 FPGA board with only 4.9Mb on-chip RAM. Since the ultra-small network parameter, it is feasible to store the whole network parameter into on-chip RAM, which could greatly reduce the energy and delay overhead to load network parameter from off-chip memory. Meanwhile, a new data streaming pipeline architecture is proposed in PCâ€“BNN FPGA implementation to further improve throughput. The experiment results show that our PCBNN based FPGA implementation achieves 930 frames per second and 387.5 FPS/Watt, which are among the best throughput and energy efficiency compared to most recent works.