The increasing complexity of Convolutional Neural Networks (CNNs) and their deployment in resource-constrained environments have emphasized the importance of efficient hardware accelerators. Field Programmable Gate Arrays (FPGAs) provide an attractive balance of performance, energy efficiency, and flexibility for deep learning inference. However, implementing modern CNNs, particularly those with architectural features such as skip connections, on FPGAs presents considerable challenges related to performance, latency and resource utilization.

This thesis presents NN2FPGA, a novel high-level synthesis-based framework that automatically compiles quantized CNN models into highly efficient static dataflow accelerators targeting embedded FPGAs. The framework performs graph level optimizations for managing skip connections, provides a templated C++ library supporting various CNN layers and data types, and a binary integer programming (BIP) based design space exploration strategy that balances throughput and hardware resource usage.

A key contribution is the introduction of buffering-aware optimizations that minimize on-chip memory requirements for residual blocks through techniques such as temporal reuse and loop merging. These enable efficient fusion of operations across branches while preserving dataflow parallelism. Furthermore, the BIP formulation enables optimal selection of loop unroll factors across layers, maximizing throughput under DSP and memory constraints and subsequently minimizing resource usage for fixed performance targets.

The proposed framework is validated on several state-of-the-art CNN models, including ResNet8, ResNet20, and MobileNetV2, using datasets such as CIFAR-10 and ImageNet, and deployed on various Xilinx/AMD FPGA platforms. Experimental results demonstrate throughput improvements of up to 7× compared to other HLS based tools and even a 10% speedup over a handcrafted RTL design, confirming the viability of HLS-based CNN acceleration when guided by rigorous optimization.

This thesis proposes nn2FPGA as a scalable and robust methodology for hardware aware CNN compilation and provides the basis for future research to support more neural network topologies and application domains, including object detection and semantic segmentation.