Title	FPGA向けのオフチップメモリ参照を削減したブロックベースCNNアクセラレータに関する研究
Author(s)	陳, 炎
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19924
Rights	
Description	Supervisor: 田中 清史, 先端科学技術研究科, 博士

Abstract

The rapid advancement of convolutional neural networks (CNNs) has revolutionized various fields since AlexNet's success in 2012. While GPUs excel at training CNNs, hardware accelerators are vital for inference, leveraging 8-bit or lower quantized data for energy efficiency. However, challenges remain with memory access and computational efficiency. This research introduces a novel CNN accelerator to address these issues.

CNN accelerators generally fall into Overlay and Dataflow categories. Overlay designs sequentially process layers with a unified Processing Element (PE) Array, offering flexibility but heavily relying on off-chip memory, which limits efficiency for lightweight models like MobileNetV2. Dataflow designs dedicate hardware to each layer, minimizing memory access but requiring extensive on-chip memory to store weights, reducing flexibility.

The proposed block-based architecture leverages the strengths of existing designs while addressing their limitations, enabling tailored optimization for different network structures. It incorporates multiple PE Arrays and supports flexible runtime interconnection modes: serial execution, which accelerates entire blocks, and parallel execution, which processes individual layers.

Experiments on 7Z010, ZU3, ZU7, and VU13P FPGAs show significant performance gains, achieving up to 11,821 FPS on VU13P with 8-bit quantized MobileNetV2. Despite having a minimal area requirement comparable to typical Overlay designs, it achieves significantly higher throughput per unit area than state-of-the-art accelerators. Compared to typical Overlay designs, our design reduces overall off-chip memory transfer volume by 93%; compared to typical Dataflow designs, our design reduces the on-chip weight storage requirement by 88%, offering a scalable, high-performance solution for modern CNNs.

The minor research optimizes the accelerators by balancing PE Array size, memory, and DSP allocation. Simulated Annealing (SA) and Genetic Algorithm (GA) are explored, with SA providing more stable results. Experiments on various FPGAs demonstrate significant throughput and area efficiency improvements over manual configurations, highlighting SA's practical utility.

Keywords: FPGA, Hardware Accelerators, Convolutional Neural Networks, MobileNetV2, YOLOv3.