

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

氏 名	CHEN, Yan		
学 位 の 種 類	博士（情報科学）		
学 位 記 番 号	博情第 546 号		
学 位 授 与 年 月 日	令和 7 年 3 月 21 日		
論 文 題 目	A Study on Block-based CNN Accelerators with Reduced Off-chip Memory References for FPGAs		
論 文 審 査 委 員	田中 清史	北陸先端科学技術大学院大学	教授
	井口 寧	同	教授
	丹 康雄	同	教授
	リム 勇仁	同	准教授
	松谷 宏紀	慶應義塾大学	教授

論文の内容の要旨

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.

論文審査の結果の要旨

本博士論文は、畳み込みニューラルネットワーク（CNN）の計算需要の増大に伴う課題に対応するため、新しい FPGA ベースのアクセラレータアーキテクチャを提案している。高いエネルギー効率を実現するために、低精度データで推論を行うアクセラレータが不可欠であるが、メモリアクセスと計算効率が重要な課題となっている。既存のオーバーレイアーキテクチャは、異なるレイヤーの計算タスクを統一された PE アレイに割り当てることで柔軟性を提供するが、オフチップメモリへの大きな依存が性能を制約し、このことが特に計算量が小さい軽量なネットワークでは問題となる。一方、既存のデータフローアーキテクチャは、レイヤーを専用モジュールにマッピングし、それらを直列に接続することでメモリアクセスを軽減するが、大量のオンチップメモリを必要とするため、柔軟性に制限がある。

本博士論文は、オーバーレイ設計の柔軟性とデータフロー設計の効率性を組み合わせた新しいブロックベースのアクセラレータアーキテクチャを提案している。これには、複数の PE アレイと、直列および並列実行モードの両方をサポートするランタイム相互接続メカニズムが含まれる。直列実行は、複数の PE アレイを直列に接続し、ブロック全体の処理を高速化する。一方、並列実行は複数の PE アレイを並列に接続し、各レイヤーの処理を並列化により高速化する。この二重モードの運用により、さまざまなネットワーク構造に最適なりソース利用を実現する。この設計における重要な革新点は、オフチップメモリへの依存を削減することにある。レイヤー間のデータの一時的な保存を不要にするため、アーキテクチャ内に相互接続バスを組み込み、さらに、十分なオンチップメモリを持つデバイスでは、ブロック間データのバッファリングを行うことでオフチップメモリへの転送を削減し、キャッシュ無しで最大 88.9%、バッファリングを用いることで 100%の削減を達成している。

本研究では、様々な FPGA プラットフォーム上で実験を行い、設計のスケーラビリティと性能を評価した。MobileNetV2 を実行した際、提案するアクセラレータは VU13P FPGA 上で最大 11,821 FPS を達成し、最先端のアクセラレータのスループットを大幅に上回る性能を示した。さらに、このアーキテクチャは従来のオーバーレイ設計と同等の回路面積で実装可能でありながら、単位面積あたりのスループットにおいてより優れた性能を発揮している。本研究における実装は MobileNetV2 のスループットが初めて 10,000 FPS を超えたものであり、今後のアクセラレータ設計の新たなベンチマークを確立するとともに、他のアーキテクチャを評価するための基準となることが予想される。

以上、本論文は、深層学習推論アクセラレータの大幅な性能向上を達成するアーキテクチャを提案・評価したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。