論文 / 著書情報
Article / Book Information

題目(和文)	FPGAベースの機械学習アクセラレータの設計最適化に関する研究
Title(English)	A Study on Design Optimization for FPGA-based Machine Learning Accelerator
著者(和文)	神宮司明良
Author(English)	Akira Jinguji
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11761号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:中原 啓貴,髙橋 篤司,本村 真人,劉 載勲,佐々木 広,高前田 伸也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11761号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		神宮司　明良	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	中原　啓貴	准教授	審査員	佐々木　広	准教授
	審査員	高橋　篤司	教授		高前田　伸也	准教授
		本村　真人	教授			
		劉　載勲	准教授			

論文審査の要旨（2000字程度）

博士論文「A Study on Design Optimization for FPGA-based Machine Learning Accelerator」（日訳：FPGAベースの機械学習アクセラレータの設計最適化に関する研究）は５章から構成されており、FPGAを設計対象とした機械学習向けアクセラレータの設計方法、並びにフレームワークを提案した。

第一章では自動化・無人化社会の到来について、内閣府のデータを元に提示し、自動化の実現に必要な要素の一つとして認識システムを挙げ、機械学習による認識タスクに基づく設計法の必要性を述べている。ロボット、自動運転、防犯カメラ等の自律型システムにおいてリアルタイム処理・低消費電力・安価な製造コストを達成しつつ、設計を容易にする機械学習アクセラレータを博士論文の高見点として纏めている。本論文では柔軟性・性能のバランスに優れた FPGA(Field-Programmable Gate Array)を設計対象としている。

第二章では、機械学習の一種であるランダムフォレストを題材とし、高位合成（HLS: High Level Synthesis）を用いたアクセラレータの設計法を提案している。ランダムフォレストは対象とするデータセットに応じて柔軟に変更が必要なため、従来のRTL（Register Transfer Level)設計では設計時間が長く実用に適しないため、アルゴリズムレベルでアクセラレータを設計できるHLSに着目し、ランダムフォレストをHLS向け中間表現に変換するツールkrangeを開発し、比較器をk-means法を用いて共有化し面積削減した。FPGA実装はCPUより8.4倍高速であり、クラスタリングにより認識精度とハードウェア量を自由に設計できる手法を開発した。

第三章では、画像認識処理で広く用いられている深層学習の一種である畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）に着目し、重みパラメータのスパース性を適用したアクセラレータを提案した。実アプリケーションとしてCNNベースのOpenPoseに対して、間接メモリ参照方式を提案し、FPGA実装によりGPUと比較して3.5倍の高速化、13倍の電力性能効率を達成した。スパースCNNとFPGA回路の協調設計により、高い性能を達成できることを明らかにした。

第四章では、第三章とは対象的に、特徴マップスパース化を提案した。CNNの特徴マップを分割し、時分割実行するSplit-CNNを対象とし、空間分割畳み込みに基づくメモリアクセスアーキテクチャを提案した。また、クラス分類タスクに対して空間分割数と認識精度劣化のトレードオフを明らかにした。設計したSplit-CNNアクセラレータはGPUと比較して3倍スループットを改善し、電力効率を9倍改善した。重みパラメータと特徴マップのメモリ削減手法を明らかにした。

第五章では、第四章で提案したアクセラレータを改善し、並列実行するアーキテクチャを提案した。多くの場合、畳込み演算がCNNの計算の90%以上を占める。従って、畳み込みをマルチコアで並列実行できれば高速化を達成できる。申請者は畳込み演算の入力依存性に着目し、リングバスと畳み込み演算で依存性を隠蔽するアーキテクチャを提案した。また、大規模な並列演算カーネルを小規模な演算カーネルの並列設計に変更し、HLSの制御回路を分割実装する設計を行なった。その結果、制御回路の制御信号を単純かつ短くでき、高動作周波数を実装できた。設計したアーキテクチャに開発したメモリ削減手法を適用するWasabiフレームワークを開発し、誰でもPGAを設計対象とした機械学習向けアクセラレータの最適な設計法を利用できるようにした。

以上を要するに、本論文は，重みパラメータと特徴マップ用メモリ量を削減することで電力性能効率を改善するとともに、フレームワークにより柔軟な設計を可能にしたものであり、工学上、工業上貢献するところが大きい。よって我々は本論文が博士(工学)の学位論文として十分価値あるものと認める。

注意：「論文審査の要旨及び審査員」は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。