

論文 / 著書情報
Article / Book Information

題目(和文)	RISC-VをベースとしたAI画像認識システムのための特定用途向け命令セットプロセッサ
Title(English)	Application-Specific Instruction-set Processor for AI Visual System based on RISC-V
著者(和文)	WANGHansen
Author(English)	Hansen Wang
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12861号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:一色 剛,高橋 篤司,本村 真人,佐々木 広,藤木 大地
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12861号, Conferred date:2024/9/20, 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		Wang Hansen	
論文審査 審査員		氏 名		職 名		氏 名	職 名
	主査	一色 剛		教授		藤木 大地	准教授
	審査員	高橋 篤司		教授	審査員		
		本村 真人		教授			
		佐々木 広		准教授			

論文審査の要旨（2000 字程度）

本論文は、Application-Specific Instruction-set Processor for AI Visual System based on RISC-V（RISC-V をベースとした AI 画像認識システムのための特定用途向け命令セットプロセッサ）と題し、英文 6 章から構成されている。 第一章 Introduction（緒言）では、AI (Artificial Intelligence)を用いた画像認識システムが幅広い分野で応用開発が急速に進む中、これら AI 画像認識システムを実現するハードウェアアーキテクチャの技術課題について述べている。その中で、日々進化を遂げている最新の AI アルゴリズムを利用するために、CUDA (Compute Unified Device Architecture)プログラミング環境を提供し、膨大な演算処理性能を発揮する GPU(Graphic Processing Unit)が AI 画像認識システムにおいて最も普及しているが、エッジ端末側で AI 画像認識機能を実現するためには、GPU の膨大な消費電力が大きな障害であるとしている。また、高い演算処理性能と低消費電力を両立するための ASIC (Application-Specific Integrated Circuits)方式は、その固定化されたハードウェアアーキテクチャがもたらす機能柔軟性の低さが、進化する AI アルゴリズム改良実装を困難にし、さらに、多様なアプリケーション開発を複雑にしていると指摘している。このような背景の下で、本論文は、高い機能柔軟性、高い演算処理性能、及び低い消費電力を同時に達成する AI 画像認識システムを実現するために、RISC-V 拡張命令駆動型アクセラレータアーキテクチャを新たに提案し、アクセラレータが搭載する高並列演算ユニットを高い処理稼働率で動作させるための広帯域オンチップメモリ構成と、アクセラレータの演算動作やデータ転送動作を制御する拡張命令方式によって、多様な画像認識アプリケーションが実装可能であることを示し、提案アーキテクチャの処理電力効率の優位性を既存アーキテクチャとの比較評価を通じて実証することが本論文の目的であると述べている。 第二章 Related Works（関連研究）では、AI 画像認識システムにおける既存アーキテクチャを概観し、提案アーキテクチャの開発に関連する RISC-V 命令セットと C2RTL 高位設計検証環境について解説している。その中で、1000 基以上の AI 処理に特化したプロセッサを搭載したメニーコアアーキテクチャ事例や、膨大な数の演算器とオンチップメモリを半導体基板全体に搭載したウェハースケールアーキテクチャ事例などを紹介し、膨大な演算性能を実現するためのプロセッサ構成、広帯域・大容量オンチップメモリ構成、並びにこれらの集積化アーキテクチャの技術的アプローチを整理している。 第三章 Deep Neural Network Accelerator System-on-Chip（深層ニューラルネットワークアクセラレータシステム・オン・チップ）では、深層ニューラルネットワーク処理の主要部分である膨大な積和演算性能を担保するための広帯域オンチップメモリ構成の重要性に着目し、高並列積和演算ユニットとオンチップメモリサブシステムを RISC-V プロセッサの拡張命令によって制御する機構を特徴とする新たな AI アクセラレータアーキテクチャを提案している。その中では、深層ニューラルネットワーク処理負荷の大部分を占める 2 次元畳込み演算を実現する 1152 基の積和演算アレイユニットと、これらを 100%に近い稼働率で動作させるためのラインバッファ及びシフトレジスタからなる広帯域メモリサブシステムで構成される高スループットデータパス機構を導入し、このデータパスの演算動作を詳細に制御するための拡張命令によって、異なるニューラルネットワーク構成（ResNet、VGG など）をソフトウェアで実現できることを示している。既存アーキテクチャとの処理電力効率の比較では、提案アーキテクチャが FPGA 実装や ASIC 実装に比べ 9.8 倍から 160 倍高く、さらに GPU に比べて 1893 倍高いことが示されており、提案アーキテクチャが処理電力効率において明確な優位性を示している。 第四章 Artificial Intelligence Visual System-on-Chip（人工知能画像認識システム・オン・チップ）では、前章で提案した AI アクセラレータアーキテクチャが対象とする標準的なオブジェクト認識処理だけでなく、画像領域分割や姿勢推定など、より高度な画像認識処理を実現する最新の YOLO アルゴリズムを高速実行するための改良型 AI アクセラレータアーキテクチャを提案している。その中では、前章で提案し
--

た高スループットデータパス機構を踏襲しながら、最新 YOLO アルゴリズムで使用する複雑な活性化関数を区間線形関数で近似計算する専用演算回路を導入した他、転置畳込み演算や行列乗算などに対応するためのデータ再配置機構を広帯域メモリサブシステムに搭載することで、複雑な処理構造を持つ YOLO アルゴリズム全体を改良型 AI アクセラレータ処理と RISC-V 上のソフトウェア処理の組合せで実現できることが示されている。既存アーキテクチャとの処理電力効率の比較では、提案アーキテクチャが FPGA 実装や ASIC 実装に比べ 3.1 倍から 38 倍高く、さらに GPU に比べて 166 倍高いことが示されており、複雑な最新 YOLO アルゴリズムにおいても提案アーキテクチャが処理電力効率において明確な優位性を示している。

第五章 AI Application Software Design Using RISC-V Custom Instructions (RISC-V 拡張命令による AI アプリケーションソフトウェア設計) では、本論で提案する 2 種類の AI アクセラレータアーキテクチャで実行する AI 画像認識アプリケーションのプログラミング手法について述べている。その中で、標準的なオブジェクト認識処理のための畳込み層を例に、拡張命令記述と通常プログラム記述が混載したコーディング例が示されている他、YOLO アルゴリズムに含まれる転置畳込み、行列乗算、及び Softmax 活性化関数のコーディング例についても示されており、拡張命令を含んだソフトウェア記述によって様々な AI 処理アプリケーションを容易に開発することができることが示されている。

第六章 Conclusions and Future Works (結言と今後の課題) では、本論文で得られた結果をまとめ、残された課題を述べている。

以上を要するに、本論文は、AI 画像認識機能をエッジ端末側で実現する上で、機能柔軟性と処理電力効率の両立が大きな技術課題であることに対し、アクセラレータが搭載する高並列演算ユニットを高い処理稼働率で動作させるための広帯域オンチップメモリ構成と、アクセラレータの演算動作やデータ転送動作を制御する拡張命令方式を特徴とする RISC-V 拡張命令駆動型アクセラレータアーキテクチャを新たに提案し、既存アーキテクチャと比べて処理電力効率が高いことを複数の AI 画像認識アプリケーションを通じて実証したものであり、工学上および学術上貢献することが大きい。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があるものと認める。

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。