論文 / 著書情報
Article / Book Information

題目(和文)	アーキテクチャ－アルゴリズム協創による小型・高効率ニューラルネットワークアクセラレータの研究
Title(English)	A Study of Highly Compact and Efficient Neural Network Accelerators through Architecture/Algorithm Co-Exploration
著者(和文)	安藤洸太
Author(English)	Kota Ando
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11941号, 授与年月日:2021年3月26日, 学位の種別:課程博士, 審査員:本村 真人,髙橋 篤司,劉 載勲,中原 啓貴,原 祐子,佐々木 広,高前田 伸也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11941号, Conferred date:2021/3/26, 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		安藤　洸太	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	本村 真人	教授	審査員	原 祐子	准教授
	審査員	高橋 篤司	教授		佐々木 広	准教授
		劉 載勲	准教授		高前田 伸也	准教授(東大)
		中原 啓貴	准教授			

論文審査の要旨（2000字程度）

本論文は、A study of highly compact and efficient neural network accelerators through architecture-algorithm co-exploration（アーキテクチャ-アルゴリズム協創による小型・高効率ニューラルネットワークアクセラレータの研究）と題し、英文６章から構成されている。

第１章 Introduction （序論）では急速に発展の進むニューラルネットワーク関連技術の情勢を述べている。特にこの発展を可能としたハードウェア技術としてGPUに始まる大規模な並列処理と、FPGAおよびASICによるニューラルネットワークプロセッサの概観を述べる。また広くニューラルネットワークが利用される所以である高精度を実現したアルゴリズム・モデルアーキテクチャのトレンドを述べている。応用先によって大きく要求の異なるニューラルネットワーク処理をユーザサイド・エッジサイドで高速・高効率に処理するためのアーキテクチャと、そのためのアルゴリズムの要件に関して考察を与え、以降の提案を導いている。論文の貢献として、計算処理の構造が大きく異なる畳込み層と全結合層の処理を統一したデータフロー再構成型アーキテクチャの提案とその上でのアーキテクチャ探索の実例の提示、電力を消費する外部メモリを排してSRAM直下での大規模並列処理を行うことで高効率化を図るニアメモリのニューラルネットワークプロセッサの提案、コンパクトなハードウェアで効率と精度の両立を目指す量子化アルゴリズムの提案を通し、ハードウェアとソフトウェアの両技術を俯瞰する協調設計の効果を実証すると述べている。

第２章 Background （研究背景）にて本論文の着想をなした先行研究や関連技術、並びに昨今の関連分野の研究で重要視されている周辺技術について解説している。

第３章 Architecture Exploration Focusing on the Diversity of Convolutional Neural Network Processing（畳込みニューラルネットワーク処理の多様性に着目したアーキテクチャ探索）において、畳込み層と全結合層の処理に共通する並列性を抜き出し、これを単一のアーキテクチャで処理する方法に関して考察・提案している。１対多の入力の並列処理を考えると、必要になるデータの供給パタンの差のみで双方を処理できることから、データフローのみの組み替えによりその両方を効率的に処理できることを示している。さらにこれから出発し、メモリアクセスパタンとデータ要求量、ハードウェアリソースを定量的に解析し、その最適化を行う手法について提示し、実践している。

第４章 BRein Memory: Near-Memory Processor for Quantized Neural Networks （BRein Memory: 量子化ニューラルネットワークのためのニアメモリプロセッサ）では、前章で観察したデータ転送量と効率の観点から、全ての外部メモリアクセスを排してオンチップSRAM直近での並列処理を行うニューラルネットワークプロセッサを提案している。ニューラルネットワーク各層の１入力多出力と多入力１出力という対称な並列性を抜き出しSRAM直下でビット単位の並列処理に閉じることにより、多ビットのデータ転送を完全に省く。この実現のため量子化ニューラルネットワークアルゴリズムを取り入れている。さらにLSI実装で実チップ評価を行い、量子化ニューラルネットワークのニアメモリ処理の高い電力効率を実証している（2.5 Mb SRAM・13層ネットワークを搭載可能で電力効率 2.1 TOPS/W・処理性能 1.4 TOPS)。

第５章 Dither NN: Accurate and Efficient Quantization Algorithm Enabled by Hardware-Software Co-Designing （Dither NN: ハードウェア・ソフトウェア協調による高効率・高精度量子化アルゴリズム）においては、第４章で用いたような量子化アルゴリズムで発生してしまう認識精度劣化を、ハードウェア効率を損なわずに改善するアルゴリズムとそのプロトタイプアーキテクチャを提案・評価している。ここでは、量子化ニューラルネットワークの活性化関数・量子化処理を信号処理の文脈で捉え、画像の低ビット精度量子化で用いられるディザを取り入れることを発想している。誤差拡散法ディザが加減算のみで実現できることから、ニューラルネットワークプロセッサの積和演算のための加算器をディザに転用することで、量子化アルゴリズムを採用する動機であった軽量なハードウェア

実現をそのままに認識精度を向上させることができると述べている。このアルゴリズムの学習方法の実験的検証・提案と、各種ハードウェアにおける実用に際しての最適化に関しても述べている。実際に学習と FPGA プロトタイピングを通して低ハードウェアコストでの認識精度向上を実証している。

以上を要するに、本論文はニューラルネットワークのエッジデバイスにおける活用を推進するため、アルゴリズムの並列性およびデータフローに着目したハードウェア構築と、ハードウェアの構造の要求や制約を反映したニューラルネットワークモデル・アルゴリズムの構成という両面からのアーキテクチャ–アルゴリズム協調設計の手法と実例を提示したものであり、学術的および工学的貢献は大きい。よって、審査員は本論文が博士（工学）の学位論文として十分に価値があるものと認める。