論文 / 著書情報
Article / Book Information

題目(和文)	環境適応型エッジAIの実現に向けた低電力深層学習アクセラレータ
Title(English)	Low-Power Deep Learning Accelerator for Environment-Adaptive Edge AI
著者(和文)	鈴木淳之介
Author(English)	Junnosuke Suzuki
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第286号, 授与年月日:2025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,CHU VAN THIEM,ISLAM A K M MAHFUZUL,佐々木 広
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第286号, Conferred date:2025/3/26, 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		鈴木 淳之介	
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	佐々木 広	准教授
	審査員	高橋 篤司	教授			
		Chu Van Thiem	准教授			
		ISLAM Mahfuzul	准教授			

論文審査の要旨（2000 字程度）

本論文は、「Low Power Deep Learning Accelerator for Environment-Adaptive Edge AI（環境適応型エッジ AI の実現に向けた低電力深層学習アクセラレータ）」と題し、英文 6 章から構成される。

第 1 章 Introduction（序論）では、本研究で取り組む環境適応型エッジ AI の実現に向けた課題が紹介されている。近年、エッジ AI は、クラウドの中央集権的な計算・通信負荷を軽減し、リアルタイム性やプライバシー保護が求められる分野で重要性が増している。一方で、エッジ AI は、端末の厳しいリソース制約に加え、特定タスク向けに最適化された固定的なモデルを使用しているため、動的な環境への適応が困難であるという課題がある。そのため、環境適応型エッジ AI の実現には、以下の要件が求められる。1）低電力性：エッジ端末の厳しい計算資源、電力制約を満たす。2）適応性：時間的・空間的なリソース変動に対応し、計算負荷を柔軟に調整する。3）効率性：限られた計算資源の中で、計算量―精度のトレードオフを最適化する。本研究では、これらの課題を解決するために、単一モデルでネットワーク改変を必要とせず、低電力・高効率・柔軟な適応を可能にするフレームワークを開発することを目的とする。

第 2 章 Background（研究背景）では、この論文の基盤となる基本的な研究について紹介されている。具体的には、適応的推論、量子化ニューラルネットワーク、スケーラブルなビットレベル演算器とデータパス、およびドメイン特化型 ML アクセラレータについて解説されている。

第 3 章 ProgressiveNN: Adaptive Quantization Framework With Progressive Bit-Precision（ProgressiveNN：漸進的ビット精度を用いた適応的な量子化フレームワーク）では、単一の重みセットを用いて適応的なビット精度を実現する適応的推論アルゴリズム「ProgressiveNN」が提案されている。このアルゴリズムでは、ビット単位の 2 値量子化と最上位ビット優先の累積機構を組み合わせることで、効率かつ柔軟な可変精度を実現する。また、得られた可変ビット精度モデルは、全体の 0.03%に相当するバッチ正規化層のみを再学習することにより性能が大幅に改善し、単一モデルで精度―計算量のトレードオフ調整を実現する。エントロピーしきい値を用いた信頼度ベースの動的ビット精度調整は、推論精度―計算量のトレードオフを改善し、平均ビット長 2 ビットにおいて、1.3%の精度向上を報告している。

第 4 章 Pianissimo: A Sub-mW Class DNN Accelerator With Progressive Bit-Serial Datapath（Pianissimo:漸進的なビットシリアルデータパスを用いたサブ mW クラス深層ニューラルネットワークアクセラレータ）では、第 3 章で提案したビット漸進型ニューラルネットワークを効率的に処理する、柔軟かつ超低電力推論アクセラレータが提案されている。このアクセラレータは、シンプルなビットシリアル演算器、RISC と HW カウンタの協調制御、およびユニファイドメモリを含む 3 層のメモリ階層により、サブ mW クラスの超低電力動作と適応精度による柔軟性をサポートしている。また、適応精度に加え、低電力イベント駆動型イメージセンサから得られる動体情報を活用することで、計算効率を著しく向上させることを示しており、3.0 TOPS/W から 27.7 TOPS/W の競争力のある電力効率を達成している。

第 5 章 Sparsity-Aware Progressive Bit-Precision Networks（スパース性を考慮した漸進的ビット精度ネットワーク）では、第 3 章で紹介した適応精度量子化の拡張アルゴリズムが提案されている。ビット単位 2 値量子化のゼロ表現の欠如がゼロ演算スキップによる効率化利用を妨げているという課題に対し、ブース符号化によるビット分解と段階的な累積が、単一重みを用いた適応精度とスパース性の両立を実現可能であることを示している。また、三値分を事前に学習して固定することで、特に低ビットでの性能を大幅に改善する漸進的ビット精度量子化モデルの学習手法が提案されている。提案手法は、8 ビットモデルでは精度劣化を 1%以下に抑えつつ、最適な三値モデルを内包した柔軟なモデルを提供している。

最後に、第6章 Conclusionでは、本研究の意義と貢献について総括している。本研究では、第1章で述べたエッジAIの抱える低電力性、効率性、適応性という課題に対し、アルゴリズムとハードウェア設計の協調による包括的な解決策を提示した。本論文で提示した各種の提案は、それぞれが適応精度、超低電力推論、スパース性活用の新たな可能性を示し、これらを基盤としたアプローチが、従来の固定的なエッジAIを超える柔軟で効率的なシステムの実現に貢献することが述べられている。

以上を要するに、本論文は、エッジAIが直面する低電力性、効率性、適応性の課題に対し、適応精度の動的ビット精度調整や超低電力推論アクセラレータの設計といった革新的なアルゴリズムとハードウェア設計の協調によって、環境適応型エッジAIの実現可能性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。