学位論文審査の要旨

博士の専攻分野の名称　博士 (工学)　　氏名　金子　竜也

審査担当者　主　査　教　授　池辺 将之
副　査　教　授　葛西 誠也
副　査　教　授　富田 章久
副　査　教　授　本久 順一

学位論文題名

エッジ AI ハードウェアに向けたニューラルネットワークの新学習アルゴリズムとその低電力アーキテクチャに関する研究
（Novel Neural Network Training Algorithms and their Energy Efficient Architectures for Edge-AI Hardware）

本研究は、人工知能 (AI) の学習に用いられる誤差逆伝播法と最適化手法、特にエッジ端末で学習を行うための軽量化アルゴリズムとそのアーキテクチャに関するものである。

現在、AI は画像認識、翻訳、画像生成等、種々のタスクにおいて、従来の手法に対して優位性を示している。これらの情報処理は、エッジ端末などから収集された大量のデータを基に、処理能力の高い AI エンジンを持つデータセンタ等において集中的に行われている。AI の性能 (認識能力等) は、そのモデルの複雑度およびデータ量と密接な関係を持ち、AI の高性能化とともに膨大となる演算処理を効率的に行うためのアルゴリズムやアーキテクチャ研究の必要性が高まっている。一方で、我々の身の回り (エッジ) で AI を活用することを考えた場合、データセンタでの処理に偏重した AI システムでは、セキュリティ (機密・個人情報等を含むデータがクラウドを流れてしまう)、リアルタイム性 (エッジでの電力制約によりクラウドへの常時接続は困難であるため、間欠動作による遅延が発生する)、通信帯域 (AI の学習には大量のデータが必要であり通信帯域を圧迫する)、といった問題が予測される。これらの問題に対応するために、クラウド上の AI に大きく依存せずに、ユーザの身近なところで (可能な限りオフライン環境下で)AI 処理を行う「エッジ AI」の実現が期待されている。

AI の演算処理は「推論処理」と「学習処理」に大別でき、AI の恩恵を得るためにはどちらの処理も必須である。そのうち、推論処理 (主に積和演算、畳み込み演算と非線形関数による活性化処理) については多くのエッジ AI ソリューションが存在する。一方、エッジでの学習処理 (誤差逆伝播法およびパラメータの最適化手法) については、未だ汎用プロセッサを中心としたソフトウェア的な手法に留まっており、学習処理のコスト (学習にかかる時間、消費電力、ハードウェア資源) は極めて高い。そこで本研究では、電力やハードウェア資源が制限されるエッジにおいて、学習処理を可能とする新規アルゴリズムとそれらを実装するアーキテクチャの構築を目的とした。

最初に、ニューラルネットワークの最も基本的な学習方法である誤差逆伝播法と確率的勾配降下法のハードウェア指向アルゴリズムを構築した。推論処理のみを行う従来のエッジ AI 研究では、「推論処理」の演算における数値表現方式を浮動小数点方式から固定小数点方式へと変更、または固定小数点方式のビット数を削減することで、演算の軽量化を実現してきた。本研究では、推論処理のみならず「学習処理」における演算のビット精度を制限する (固定小数点方式とする) ことで演算を軽量化した。学習処理を行う演算アーキテクチャを構築して評価した結果、性能を維持するために必要となる最低ビット数、および提案アーキテクチャの並列度を可変にすることで広範な用途に対応できることを明らかにした。

次に、誤差逆伝播法を軽量化する手法を提案し、そのアーキテクチャ構築を行った。上述の固定小数点方式の導入により演算を軽量化した結果、推論処理と学習処理では要求される最低ビット数が異なり、学習時に AI 性能に影響を与えるまで変化するパラメータ数は極小数であることを明らかにした。このことは、パラメータを保存するメモリ容量や、パラメータの更新 (メモリへの書き込み) に必要な電力等の大部分は無為に消費されていることを意味する。この発見に基づき、AI 性能を維持したまま、誤差逆伝播法のビット数 (メモリ容量)、およびメモリへの書き込み回数を削減可能な新アルゴリズムを構築した。さらに、演算に必要なハードウェア資源量が少ないエッジ AI 向け学習アーキテクチャを構築し FPGA に実装して評価した。既存手法と比較して、使用メモリ量を 49.8% 削減可能であること、およびメモリアクセスに係る消費電力を 0.0017 倍に削減可能であることを示した。

上述の研究・成果はディジタル回路を対象としたものであるが、次の挑戦として、アナログ回路を導入した「コンピューティングインメモリ (Computing in Memory: CIM) デバイス」のための誤差逆伝播法、および CIM デバイスを用いて学習処理を行うアーキテクチャの構築を行った。AI は多量の積和演算を必要とするため、従来のノイマン型の計算機で AI 演算を行うと、プロセッサとメモリ間のデータ転送に係るボトルネックにより演算に膨大な時間がかかってしまう。この問題の解消に向けた取り組みとして多量のデータ (パラメータ) を保存しているメモリ上で積和演算を行う CIM アーキテクチャが注目されている。本研究では、非ノイマン型アーキテクチャの一種である「ReRAM を用いた CIM AI デバイス」の学習機能の実現に向けた新規アルゴリズムの開発を行った。通常の誤差逆伝播法は、活性化前のアナログ値の読み出しを必要とするため、多くの CIM AI デバイスには適さない。そこで、ディジタル誤差逆伝播法 (Digital BP) に着目し、その弱点 (性能低下) を補うニューラルネットワークの構造を新たに考案した。この新構造によって、従来の Digital BP では不可能だった線形回帰や多クラスの識別が可能になることを示した。さらに、Digital BP の演算を行うアーキテクチャ構築とその FPGA 実装を行い、演算コアの消費電力を 10 mW 以下にできることを示した。

最後に、軽量性と高性能を両立する最適化手法の開発を行った。従来のエッジ AI 向け学習アルゴリズム・アーキテクチャ研究では、最適化手法として主に確率的勾配降下法が用いられている。確率的勾配降下法は、極めて単純な最適化手法であるが故に学習の収束性や安定性が悪い。より高度な最適化手法は、大容量のメモリと高度な演算処理を必要とするため、エッジ AI 領域ではこれまで確率的勾配降下法を採用せざるを得ない状況にあった。そこで本研究では、メモリ容量と演算量を削減する新規最適化手法のアルゴリズムを構築した。ディジタル処理では一般的に精度を落とす要因となる「量子化」を積極的に利用する (固定小数点による量子化と、乗除算をビットシフトで代替できる対数量子化を組合せる) ことで、最上ランクの最適化手法 (RMSProp 等) と同程度の性能を持ちつつ、省メモリ・省リソースの最適化ハードウェアが構築可能であることを示した。高性能な最適化ハードウェアをエッジ AI に組込むことができれば、高速な学習の収束性、すなわち学習回数を削減できる。一回の学習に係るリソースを減らして低電力化するという従前のアプローチとは異なり、学習回数 (ReRAM の書き込み回数) を減らすことで低電力化する。提案手法は、従来手法と比べて約 70% の省メモリ化と 4 倍の高速化を達成可能であることを示した。

これを要するに、本研究はエッジ AI 向けの新規学習アルゴリズムおよびその回路アーキテクチャの構築法を確立したものであり、AI と半導体集積回路技術とを結びつける学際的な研究分野に対して貢献するところ大なるものがある。よって、著者は北海道大学博士 (工学) の学位を授与される資格があるものと認める。