学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（工学）　　氏名　金子　竜也

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, 人工知能技術の中でもニューラルネットワークの学習に用いられる最適化手法の, 特にエッジ端末で実行するためのアーキテクチャとアルゴリズムの両面に関するものである.

今日, 人工知能は特に機械学習のニューラルネットワークの分野において深層学習を代表に画像認識, 翻訳, 画像生成等, 種々のタスクにおいてその性能の高さを示している. これらの情報処理はエッジ端末で収集されたデータを基に, 高性能なデータセンタを用いて集中的に行う仕組みとなっている. ここで性能は人工知能モデルの複雑度と密接な関係を持ち, 高性能化と共に膨大となる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の研究の必要性が高まっており, 近年これらの研究が活発的に行われている. 一方で我々の身の回りが舞台となる実世界で人工知能を活用する場合には, クラウドであるデータセンタでの処理に偏重した AI システムに内在する (1) セキュリティ (2) リアルタイム性 (3) 通信帯域という問題が生じる. これらの問題に対応するために完全な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.

完全オフライン環境下での AI 処理にはエッジ AI 自身がその場での学習を行う必要があるため, 誤差逆伝播法と最適化手法の実装が不可欠である. そこで本研究では, 電力や演算性能が制限されるエッジ AI において演算負荷の高い学習処理を実現するためのアルゴリズム開発とそれらを実行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立的勾配降下法のハードウェア指向アルゴリズムを開発した. 従来の学習を含まないエッジ AI 研究では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することで, 演算の軽量化を実現してきた. 同様に本研究では, 数値表現に係るビット精度を固定小数点方式へと変更しビット精度を制限することで演算の軽量化を提案し, また仮想のアーキテクチャを考案した. 性能を維持するために必要となる最低限のビット精度を探査し, 提案アーキテクチャの並列性を可変にすることで広範な用途に対応できることを示した.

次に, 誤差逆伝播法を軽量化する手法を提案し, アーキテクチャの構築を行い FPGA へと実装した. ニューラルネットワークには推論と学習という二つの処理が存在する. 固定小数点という条件下ではこれらの処理で要求されるビット精度は異なり, 学習によりモデルの出力に影響を与えるまでに変化するのはパラメータの中でも極小数であることが確認できた. つまり, パラメータに関わるメモリの容量や書き込み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. 本研究では, 性能を維持したまま誤差逆伝播法のビット精度・書き込み回数を削減するアルゴリズムの提案を行った. また, 演算リソース量が最低となるエッジ AI 向けアーキテクチャを考案し FPGA へと実装した. 既存手法との比較を行い, 使用メモリ量を 49.8% 削減可能であることを示したほか, 見積もり上である

がメモリアクセスに係る消費電力を 0.0017 倍にまで削減可能であることを示した.
そして, アナログコンピューティングインメモリデバイスに向けた誤差逆伝播法の開発と, アーキテクチャの構築を行い FPGA へと実装した.
ニューラルネットワークは多量の積和演算から成り立っているために, 従来のノイマン型のディジタルアーキテクチャでは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が存在する. この問題点の解消に向けた取り組みとしてデータ (パラメータ) を保存しているメモリ上で積和演算を行うコンピューティングインメモリ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 を用いてアナログ回路的に推論処理を行う AI チップに向けたアルゴリズムの開発を行った.AI チップの特性から通常の誤差逆伝播法ではなく Digital BP 法を用い, また, より複雑なモデルに向けたアルゴリズムの改良を行った. 改良アルゴリズムは従来の手法では不可能だった線形回帰や多クラスの識別タスクの学習が可能であることを示した. また,Digital BP を処理するアーキテクチャの構築を行い FPGA へと実装し, ソフトウェアシミュレーションと同等の性能が得ることができた. 演算コア部の消費電力が 10mW 以下と低消費電力で学習可能なアーキテクチャであることを示した.
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, 最適化手法として確立的勾配降下法かこれを基にしたアルゴリズムを用いている. 極めて初歩的な最適化手法であるが故に学習の収束性や安定性等で劣るものがある. 一方で, より高度な最適化手法は慣性項を用いるためにメモリ容量の増大や, 高度な演算処理が必要となることからエッジ AI 領域では確立的勾配降下法を使用せざるを得ない状況にあった. 本研究では, 量子化を多重に施すことでメモリ容量と演算処理の削減をする最適化手法アルゴリズムの提案を行った. 高度な最適化手法を用いることは高速な収束性, すなわち学習回数の削減に繋がる.
一回の学習に係るリソースを減らすというアプローチではなく, 学習回数そのものを減らす抜本的なアプローチである. 従来手法と比べて約 70% の省メモリ化と 4 倍の高速化を両立していることを示した.