



HOKKAIDO UNIVERSITY

Title	Efficiency-Centric Hardware Accelerator for Deep Neural Network Inference [an abstract of dissertation and a summary of dissertation review]
Author(s)	植吉, 晃大
Degree Grantor	北海道大学
Degree Name	博士(工学)
Dissertation Number	甲第14129号
Issue Date	2020-03-25
Doc URL	https://hdl.handle.net/2115/78434
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Kodai_Ueyoshi_review.pdf, 審査の要旨



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 植吉 晃大

審査担当者 主査教授 浅井 哲也

副査教授 富田 章久

副査教授 葛西 誠也 (量子集積エレクトロニクス研究センター)

副査教授 池辺 将之 (量子集積エレクトロニクス研究センター)

学位論文題名

Efficiency-Centric Hardware Accelerator for Deep Neural Network Inference

(深層ニューラルネットワーク向け高効率 HW アクセラレータに関する研究)

本論文は、深層ニューラルネットワーク (DNN: DeepNeuralNetwork) の推論処理における、高効率なハードウェア実装技術に関するものである。DNN は、脳のニューラルネットワーク (NN) を模した数理モデルで、近年の人工知能技術の発展の礎となっている。特に、画像認識、音声認識、自然言語処理の分野で高い性能を達成している。これらは、大量のデータを集約して学習を行い、それらを実環境上で実行 (推論) することで、様々な応用が期待されている。そのため、実環境デバイスの厳しい制約化で推論を実行するために、省エネルギーな組み込みハードウェア技術が必須となる。本論文では、高い電力効率で DNN をハードウェア実装するために、アルゴリズムとアーキテクチャの双方から最適な方法を探索した。特に、量子化技術、アーキテクチャ探索、モデルの効率化の3点を中心に評価し、新たな手法を提案した。

まず、DNN 計算の多くを占める積和演算を軽量化するために、値の量子化手法を考案した。具体的には、値を対数領域で近似させる「対数量子化」と、実数領域で等分させる「線形量子化」の比較評価を行った。2 を底とする対数量子化により、2 進数を用いるデジタル回路上で高効率な計算を行うことができる。さらに、小さいビット幅で効率的に演算を行うアーキテクチャを考案・評価した。その結果から、ビット精度が可変な「ビットシリアル機構」を効率的に実現するアーキテクチャを考案した。このビット幅の可変メカニズムにより、精度と電力効率・速度のトレードオフを制御することが可能となった。

次いで、高い電力効率で DNN 計算を行うアーキテクチャの探索を行った。人工知能の応用技術の性能向上に伴って DNN の規模や複雑さが増大している。そのため、DNN を構築する全てのパラメータを1チップに格納することは難しくなっており、パラメータ保持のための外部メモリが必須となっている。多くのハードウェア DNN アクセラレータは外部メモリに DRAM を利用しているが、その消費電力のほとんどは外部メモリへのアクセスが占めている。本論文では、集積回路の三次元積層技術に着目し、外部メモリとして SRAM および DRAM を積層した場合の電力効率評価を行った。その結果、DNN 推論計算に許容される程度のメモリ容量であれば、SRAM を積層することで、オンパッケージ上で高い電力効率を実現できることを明かにした。さらに、これらを用いた DNN 推論計算を行うハードウェアアクセラレータを提案し、三次元積層による高バンド幅と高並列性を活かした「複数の演算コアによる並列 DNN 演算アーキテクチャ」を考案し、それを実チップ上に実装して評価を行った。

最後に、DNN モデルの効率化を行った。DNN モデルにおける計算量を削減する手法として「活性予測機構」を新に提案した。近年の多くの DNN は、その要素回路 (ニューロン) がゼロ値を出力するスパース演算となっている。本論文では、ゼロ値を出力するニューロン (無効ニューロン) をあらかじめ予測することで、計算量を削減する手法を提案した。具体的には、無効ニューロンを予測するための専用 NN を設け、学習済みの NN とは別に予測を行う手法を提案した。この予測専用 NN は二値化 NN で構成されるため、小さな面積・電力オーバーヘッドで予測を可能とする。この NN を実装・評価し、元のネットワークの不必要な計算を入力に応じて動的に予測し、計算量を削減できることを示した。

今日の人工知能技術の発展は、集積回路技術の進化とともに、大量のデータを現実時間内で処理可能となったことに起因する。これらの処理を電力制約の厳しい実世界のデバイス上で実現させるためには、ハードウェア単体技術だけでは難しく、アルゴリズムやアーキテクチャ側の見直しが非常に重要な役目を担う。これらを統合的に考慮して一つのシステムを実現させる本研究は、高度な知能情報処理のプラットフォームを拡張し、新規アプリケーション創出の可能性を示唆させる重要な役割を果たすと考えられる。

これを要するに、本研究は高効率な深層ニューラルネットワークのハードウェア実装法を確立したものであり、データサイエンス・人工知能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。