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_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

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　植吉 晃大

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: Deep Neural Network) 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、人間の脳神経 (ニューロン) 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する。ここでは、量子化技術、アーキテクチャ探索、モデルの効率化の 3 点を評価し、新たな手法を提案する。

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える。そこで、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う。2 を底とする対数量子化により、2 進数を用いるディ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計算させるアーキテクチャを評価する。そして、ビット精度を可変にすることができる、ビットシリアル機構を効率的に実現させるアーキテクチャを提案する。これにより、ビット幅を可変にすることで精度と電力効率・速度のトレードオフを制御することが可能となる。

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、オンチップ上にニューラルネットワークを構築するパラメータが全て格納することは難しい。多くの HW アクセラレータは外部メモリに DRAM を利用しているが、実際は電力のほとんどをこの外部メモリアクセスが占めることが知られている。そこで、本論文では、三次元積層技術に着目し、代表的な実用メモリである、SRAM と DRAM の評価をする。その結果、DNN 推論計算に許容される程度の容量であれば、SRAM を積層させることで、オ

ンパッケージ上で高い電力効率を実現できることを示す。さらに、これらを用いた DNN 推論計算を行う HW アクセラレータを提案する。三次元積層によって高バンド幅と高並列性が得られるため、複数の演算コアによる並列アーキテクチャを提案し、実チップ上に実装・評価をする。

最後に、DNN モデル自体の計算量を削減する手法として、活性予測機構を提案する。近年、多くの DNN は、ニューロンがゼロを出力する、スパースな演算となっている。本論文では、このゼロを出力するニューロン (無効ニューロン) を予測することで、計算量を削減する手法を提案する。無効ニューロンを予測するための専用 NN を学習済みの NN とは別に設け、独立に予測を行う手法を提案する。この予測専用 NN は、二値化された NN で構成されるため、 小さな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す。

以上、本論文に記した研究は、DNN の推論処理における、高効率なハードウェア実装を目指したものである。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、HW 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。