公開

（事後評価・研究者用）

研　究　報　告　書

「省電力なメモリアクセスを実現する DNN モデル学習」

研究期間： 2018 年 10 月～2020 年 3 月
研究者番号： 50167
研 究 者： 植吉　晃大

1.　研究のねらい

近年、深層学習(DL: Deep Learning)は画像・音声認識、自然言語処理の分野で極めて高い性能を示している。DL は深層ニューラルネットワーク(DNN: Deep Neural Network)を学習することで、予測・分類の高い性能を示す。そんな中、組み込みシステムにおいて、クラウドに頼らない計算処理の方法として、エッジコンピューティングが注目されている。エッジデバイス内部の閉じた環境にて、低遅延に高度な処理を行うことを目指している。しかし、DNN は高性能化に伴って大規模化しているため、電力・面積制約の厳しいエッジ環境下での処理は困難を極める。本研究では、こうした限られた制約下で、エネルギー効率の高い計算処理を実現するために、ハードウェア構造を意識したネットワークアルゴリズム技術の創出を目指す。

深層学習の計算はデータ量が膨大で、高いエネルギーを消費する。特に、ハードウェアにおいて、外部メモリの読み書きが、電力消費量の大半を占める。本研究では、これを最小化させることに注目する。その実現には、以下の2つの手法が考えられる。

- 計算に用いるデータ総量自体を最小化させる。
- 外部メモリアクセス頻度を最小化させる。

前者は、不必要な特徴量を間引く、枝刈り技術によってデータ総量を削減する手法が知られているが、これは汎化性を失うため、必要最小限に留めたい。そこで、本研究では後者の、外部メモリアクセス頻度を最小化させるために、新たなネットワークモデルを提案することを目的とする。学習モデルのデータ量の削減を最小限に抑え、必要な特徴量を実行時に動的に選択することで、汎化性能を失わずに、外部メモリへの不要なデータアクセス頻度を最小化させることができる。本研究では、この動的特徴選択機構を最小資源で実現できるような軽量なニューラルネットワークで実現し、トータルの電力効率を改善することを目標とする。

最終的に、これらを実行するハードウェアアクセラレータの計算機構を設計し、電力制約が厳しいエッジコンピューティングにおいて、ハードウェア・アルゴリズムの両方でのみ最適化可能なエネルギー効率と認識精度を実現する。

2.　研究成果

（1）概要

本研究では、「動的特徴選択による DNN 計算量の削減」を目標に、ハードウェアにおける計算効率化を実現するためのネットワークモデルの創出を行った。図1にその概要を示す。従来は学習モデルを圧縮縮小させることに注力を注いでいた。しかし本研究の目的は、学習した全てのパラメーターを使用せず、都度の入力に応じて、計算に必要な箇所を事前に予測することである。 その結果、計算量を必要最小限に抑えることができるとともに、メモリア

クセスを削減することも可能となる。

多くの DNN モデルでは、ニューロンの大部分が最終的に0を出力する。この計算は、次の計算結果に直接影響を与えない。この不必要な演算によって膨大なエネルギーを浪費している。この不要な計算を軽減するために、軽量な予測機構を用いて、ニューロンの活性を事前に予測する機構を提案した（研究テーマ A）。

さらに、予測機構を含めて専用回路の電力シミュレーションから、学習済みモデルを最適に実行するハードウェアアーキテクチャを提案し、評価した（研究テーマ B）。

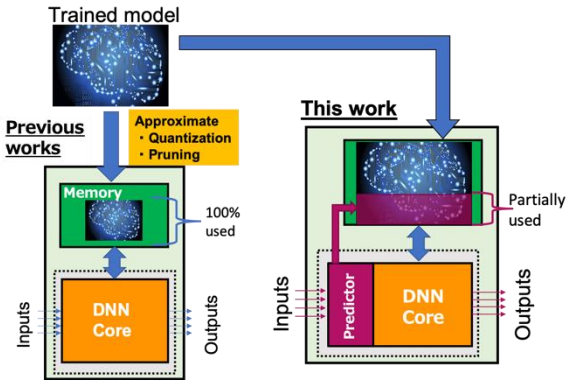

図1　本研究の概要図

（2）詳細

研究テーマ A「動的特徴選択による DNN 計算量の削減」

ここでは、DNN の出力が0となるニューロンを事前に予測する機構を評価した。多くの DNN モデルは、Rectified Linear Unit (ReLU)関数を活性化関数として用いる。これは、f(x) = min(0, x)で表され、ニューロン活性前の値が負の場合は0を出力する。図2に示す通り、一般的な画像分類のベンチマークに置いて、ニューロンの半数以上が0を出力する。本研究では、ニューロン活性前の値の正負予測として、軽量な二値化ニューラルネットワークを用いた。本予測器は、各層の入力に応じて動的に0を出力するニューロンを予測するように、層毎に独立で学習した(図 3)。

これにより、不要なニューロン計算を7割予測することができることを示した[研究成果リスト(3)-1]。また、不要なニューロン計算を予測する信頼度を調節し、認識精度低下と計算削減量のトレードオフを後から調節する手法を考案した[研究成果リスト(3)-2]。

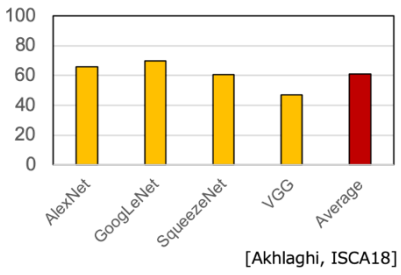

図2　画像分類(ImageNet)における
0を出力するニューロン率

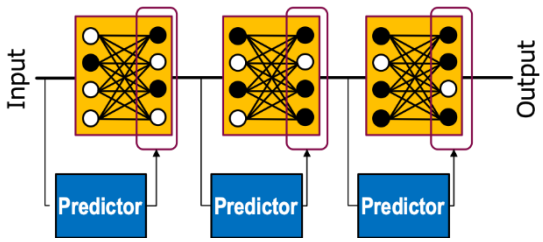

図3　提案した予測機構

研究テーマ B「動的予測型 DNN 計算アーキテクチャ」

DNN 計算とその計算予測器を効率的にパイプライン化させることにより、不必要なニューロン計算を削減させるだけではなく、電力性能・計算性能の両方を向上させることが可能である。研究テーマ A で提案した、動的な予測器を用いた計算機構では、軽量な予測器が常に動作し、DNN 専用回路のコントローラとして動作する。スパース計算に対応した DNN 専用回路を用いれば、必要最小限の計算を予測器の指示に従って行うことができる。本研究では、密な二値化ニューラルネットワークを用いた動的予測器とスパース計算に対応した多値ニューラルネットワーク専用回路を混合実装するアーキテクチャを提案した（図4）。

研究テーマ A において、事前に学習された予測器を計算する軽量なアーキテクチャとして、二値化ニューラルネットワーク計算を軽量に行うインメモリアクセラレータを採用し、実専用回路の合成ソフトを用いて電力シミュレーションを行った。その結果、通常のスパース対応した DNN 計算器に比べて、2.06 倍の電力効率を示した。

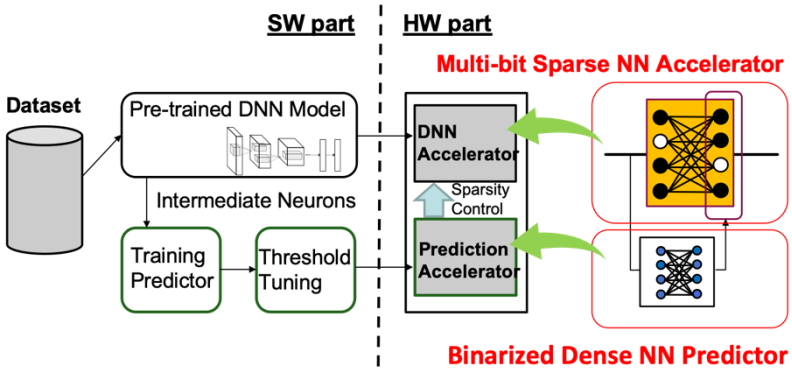

図4　動的予測器を用いた DNN 計算機構の概要

3.　今後の展開

組み込み向けの DNN プロセッサは様々な形態で研究が行われている。その中でも、ネットワークモデルの観点から、ハードウェアの最適化を行う研究は、この分野のさきがけとなっている。本研究で提案された動的予測機構は、DNN の少しのオーバーヘッドと追加学習で、様々なモデルに対応することができる。これにより、産業界においても、各利用体系に応じて、必要なニーズに合わせて、電力と性能の効率化を図った実装モデルを提案することができると考えられる。

4.　自己評価

- 研究目的の達成状況

ACT-I 期間の中で、手探り状態であったアイディアの中から、提案と結果を結びつけることができた。さらに、これらを実装フェーズで評価できたことは、当初の研究目的を達成することができた。しかし、論文の採択に結びつかず、研究期間が終わってしまったため、アウトリーチに課題が残った。

- 研究の進め方（研究実施体制及び研究費執行状況）

本研究において，発案から実装に至るまで，全て研究代表者のみで実施した。しかし、研究

室内部で常にミーティングを行い、研究室メンバや指導教員らとともにアイディアを出し合い、短い期間ながら、成果を上げることができた。指導教員らの所属で転々とした事もあったが、研究室の体制を大いに利用させていただき、指導教員の秘書らのご協力を得ながら、責任をもって研究活動を行った。

・　研究成果の科学技術及び学術・産業・社会・文化への波及効果

本研究は学術的にもホットな領域であり、最先端の成果となっている。特に、本研究では、現在利用可能なモデルに追加実装するのみで実用可能である。さらに、精度と電力性能のトレードオフを後から調節が可能であるため、産業の中ですぐに応用しやすい手法となっている。この研究が普及されることにより、厳しい制約のデバイスに搭載し得るアプリケーションの性能をさらに向上させることにつながると考えられる。

・　研究課題の独創性・挑戦性

本研究は、ハードウェアとアルゴリズムの両面から新しい改善策を見出し、1つの大きなシステムを包括的に設計することで、新しいソリューションを創出するという、非常に先進的な道を開拓している。ハードウェアの特徴を活かして、新しいモデルを構築していくという点で、一線を画す研究となっている。

5.　主な研究成果リスト

（1）論文（原著論文）発表: 0件

（2）特許出願

研究期間累積件数:1 件

1.

発　明　者:　植吉晃大、 高前田伸也
発明の名称　:　ニューラル計算装置、および、ニューラル計算方法
出　願　人:　国立大学法人　北海道大学
出 願 番 号:　特願 2019-080194

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 植吉 晃大, 池田 泰我, 安藤 洸太, 廣瀬 一俊, 浅井 哲也, 高前田 伸也, 本村 真人, "無効ニューロン予測による DNN 計算効率化手法," 電子情報通信学会リコンフィギャラブルシステム研究会, 東京 (May 2019). 優秀講演賞受賞.
2. 池田 泰我, 植吉 晃大, 安藤 洸太, 廣瀬 一俊, 浅井 哲也, 本村 真人, 高前田 伸也, "効率的な DNN 計算のための無効ニューロン予測手法の評価," 電子情報通信学会コンピュータシステム研究会, 鹿児島 (June 2019). 若手奨励賞受賞.