様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月　 ２ 日現在

機関番号：３７１１２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１８１８

研究課題名（和文）次元分離畳込み型軽量化機械学習モデルを活かすSRAMコンピューティングインメモリ

研究課題名（英文）An SRAM Computing in Memory to Exploit Energy Efficient Dimensional Separable Compact Machine Learning Model

研究代表者
山内　寛行（Yamauchi, Hiroyuki）

福岡工業大学・情報工学部・教授

研究者番号：７０４２５２３９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究はCNNモデルの無駄な計算を削減することを目指し、チャネル方向の次元削減が、１）精度を落とさずパラメータを削減できる最も効率の良い手段である、２）ハードウエア実装の点でも貢献度が少ないチャナネルを刈り取ることは比較的容易であるとの検討結果を得た。本研究では、SE-ResNet Attention　アルゴリズムを用い、実験結果からAttentionすべき特徴マップの決定プロセスにバイナリー法を用いることで、学習曲線を高速化可能であることを明らかにした。ResNetの構成を独自にコンパクトにしたResNet14構成の3ブロックで平均チャネル削減割合が約50%になることを明確にした。

研究成果の学術的意義や社会的意義
機械学習用のMemoryとComputingの二役をこなす Computing in Memory (CIM)が電力削減を目的に注目されている。しかし現状のCIM構造ではスパース領域を「電力削減制御や、逆に精度補償用の多値量子化」に活用できない。本研究は､アーキテクチャによる次元分離に欠かせないCIMを開発し、得られたスパース領域を実際の動作停止/電源遮断につなげる。省電力化に結びつかないスパース領域は多値表現に適応的に活用し2値化による精度劣化を補償するなど、精度と電力削減のトレードオフの問題を解決する。結果、時代が求めている「どこでもAIに向けての省電力機械学習エンジン」を可能にする。

研究成果の概要（英文）：This research aims to reduce unnecessary calculations in CNN models, and the dimensionality reduction in the channel direction is 1) the most efficient means of reducing parameters without reducing accuracy, and 2) also contributes in terms of hardware implementation. The study results showed that it is relatively easy to prune channels with low frequency. In this study, we used the SE-ResNet Attention algorithm and demonstrated from experimental results that it is possible to speed up the learning curve by using a binary method in the process of determining the feature map to be attended to. We clarified that the average channel reduction rate is approximately 50% in three blocks of ResNet14 configuration, which is a uniquely compact ResNet configuration.

研究分野：Computing in Memory (CIM)、どこでもAIに向けての省電力機械学習エンジン

キーワード：どこでもAIに向けての省電力機械学習エンジン　Computing in Memory

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

機械学習用の Memory と Computing の二役をこなす Computing in Memory (CIM)の回路研究が電力削減を目的に始まる一方で、モデルアーキテクチャ側の観点から次元分離によるスパース化の提案がなされている。しかし現状の CIM 構造ではその提案を原理的に実装することができず、スパース領域を「電力削減制御や、逆に精度補償用の多値量子化」に活用できない。

機械学習計算の電力消費の主要因であるデータメモリとのアクセスを削減するために、「一連の乗加算処理をビット線（BL）上で完結する SRAM 型 CIM」が 2016 年に発表され大きな反響を呼んだ [J.Zhang, N.Verma, et al, IEEE Symposium on VLSI Circuits 2016]。

しかし学習モデルの大規模化に伴う BL 上のワード数の増大で S/N 比が急激に劣化しており A/D 変換器などのアナログ補償回路依存度が高くなり電力/面積コスト課題が大きい。

原因は、1 本の BL 上での乗加算完結に固執しているからである。

その CIM 構造ではモデルアーキテクチャの新しい要求にも対応できず、次元分離による計算量削減とスパース化による細かな動作停止/電源遮断が原理的にできない。

乗加算工程の動作停止手段を細分化/階層化し、スパース情報をより反映する必要がある。

２．研究の目的

本研究の目的は CIM が抱える２つの課題：①新モデルアーキテクチャの要求(1x1 畳込み導入で次元分離)に対応できず相乗効果が発揮できない、②スパース化を CIM 上の電力削減/精度補償に活かせない。この２つの根本原因を明確にし、課題を解決することである。

本研究は、アーキテクチャによる次元分離に欠かせない 1x1 演算対応 CIM を初めて開発し、「①得られたスパース領域を CIM 回路上での実際の動作停止/電源遮断につなげる。一方で、②省電力化に結びつかないスパース領域は多値表現に適応的に活用し 2 値化による精度劣化を補償する」など、単独ではなし得ない精度と電力削減のトレードオフの問題を解決する。

結果、時代が求めている「どこでも AI に向けての省電力機械学習エンジン」を可能にする。

３．研究の方法

本研究を通じて明らかにしたい項目とその手順を、1x1 畳込み階層化 CIM コア回路設計、CIM 構造に基づく電力削減/精度補償指向スパース化/量子化協調設計、TEG 設計/試作評価の順で説明する。

【1x1 畳込みベース階層化 CIM コア回路設計】

軽量化アーキテクチャの要求は 1x1 畳込みで次元削減することである。よって、1x1 畳込みを階層化で実現する CIM 設計が可能なことを最初に明らかにする。申請者は 100 近い SRAM の階層化構造をヒントに本申請 CIM 用途に 20 程度の候補を創出している。

大事なのは 1x1 畳込み→累積和演算の一連の流れを細分化し「電源遮断/ワード線停止/AND 出力データ線プリチャージ停止」などの制御単位とスパース領域の単位をどのように「移動/スキップ/拡張」して整合させるかである。多岐にわたるが、以下の点も明らかにする。

1)　CIM 動作/電源遮断条件をより高い確率で満足するように、スパース領域を「移動/スキップ」で拡張させる一方で、CIM 制御単位自体も可変にする手段を明らかにする。

2)　スパース情報を反映した階層化制御線 C1/C2 を 1x1 畳込み用の 2-MOSFET からなる AND 回路

の GND と出力加算線のプリチャージ電源 VDD に埋め込む。その効果として、プリチャージや AND 演算の動作停止/電源遮断確率の増加量、副作用を明らかにする。

3) 動作遮断確率の観点で重み Wi と入力/活性化(IN/Act)のどちらをセル内に記憶するか？

4) スパース化の偏り（データセット/ネットワーク依存）の影響を明らかにする。

5) 1x1→1x3→3x3 への応用展開、両者混在設計手法を明らかにする。

【CIM 構造に基づく電力削減/精度補償指向量子化/スパース化協調設計】

「スパース化領域を疑似的に移動/スキップして実効的に CIM 動作単位を跨がないように拡張できること」逆に「省電力寄与率が低いスパース領域を精度補償用に多値化に活用できること」の両方を明らかにする。CIM 上で電力削減につながらないと意味がないので従来のように 0 の数でなく電力遮断/動作停止確率を報酬にして学習する手段とその効果を CIM 回路構造/データセット依存性も含めて初めて明らかにできる。

４．研究成果

初年度では、本研究は機械学習用の Memory と Computing の二役をこなす Computing in Memory (CIM)において、次元分離によるスパース化を利用した電力削減を目指す。現状の CIM 構造では原理的にスパース領域を「電力削減制御や、逆に精度補償用の多値量子化」に活用できるように実装することができない課題がある。本研究は、次 元分離に欠かせない 1x1 演算対応 CIM を開発し、得られたスパース領域を CIM 回路上での実際の動作停止/電源遮断につなげる。一方で、省電力化に結びつかないス パース領域は多値表現に適応的に活用し 2 値化による精度劣化を補償する」など、単独ではなし得ない精度と電力削減のトレードオフの問題を解決することを 狙っている。計画は、次元削減するために必要な 1x1 畳込みを階層化で実現する CIM 設計が可能なことを明らかにすることであった。それに対し当年度は、 Pytorch で Cifar10 と Cifar100 ネットワークを入力する時に必要な ResNet 型ネットワーク層に設計し、その中のボトルネック層を構成する 1×1 と 3×3 の畳込み層 を SRAM 回路で実現する場合を想定し、畳込み結果を出力する階層化ビット線の設計と制御回路アーキテクチャの設計を行った。又、ResNet20, ResNet32 のネット ワークの各層に BNN を適用した。トレーニング後のスパース化の状態を調査し、スパース化による電源遮断可能領域と物理的な回路の電源の階層化の関係を調査 した。初期データの解析の結果、ビット線や読み出し回路の階層化のパターンを何種類か用意すれば、スパース化を活かした電源遮断が可能なことが確認でき た。現在の方法を継続しデータを増やしていけば当初の計画通り、次元削減することを実現するために必要な 1x1 畳込みを階層化で実現する CIM 設計が可能なこと を明らかにすることが可能なことがわかった。

次元削減するために必要な 1x1 畳込みを階層化で実現する CIM 設計が可能なことを明らかにすることであった。その計画に対して、当該年度 は、Pytorch で ResNet で Cifar10/Cifar100 を入力する時に必要なネットワーク層を設計し、その中のボトルネック層を構成する 1×1 と 3×3 の畳込み層を SRAM で実 現するための畳込み結果を表現し出力する階層化ビット戦の回路アーキテクチャの設計を実施することができた。又、ResNet20, ResNet32 のネットワークの各層 に BNN を適用して、トレーニング後のスパース化の状態を調査した。以上のことから、スパース化による電源遮断可能領域と物理的な回路の電源の階層化の関係 を調査することができた。初期データの解析の結果、ビット線や読み出し

回路の階層化のパターンを何種類か用意すれば、スパース化を活かした電源遮断が可能なことを確認することができた。

２年目では、初年度の結果と本研究の目的を踏まえ、本研究は機械学習用の Computing in Memory (CIM)の構造と省電力化を狙った先端の機械学習モデルとの整合を狙う。1x1 ビット演算対応 CIM を開発しスパース領 域を「動作停止/電源遮断につなげる」。又、残ったスパース領域は「多値表現に適応的に活用し 2 値化による精度劣化を補償する」等、精度と電力削減のトレー ドオフの問題を解決することを狙う。当年度は先端のモデルトレンドであるアテンション機構への適用を狙い、1 年目に基礎検討し ResNet 型ネットワーク層実装用の階層化ビット線の設計と制御回路アーキテクチャのアテンション機構への有効性を検証した。表現力とパラメータ数の削減に効果的なチャネルと空間方向の アテンションブロックの検証に集中した。ResNet18 ネットワークに加えてチャネルアテンションブロックを追加した SE-ResNET18 ネットワークを Cifar10 のデータ セットで学習した場合のスパースア化の状態を SE ブロック無しの場合と比較した。アテンション値をバイナリー化した場合の精度への影響を調査した。その結果、アテンション値が 0 になる特徴マップのチャネルと空間はスパース化してもインファレンスの精度への影響は小さく、1 年目に検討した結果よりもさらにス パース化できる領域は広がる可能性が明らかになった。そのため、スパース化による電源遮断可能領域と物理的な回路の電源の階層化の関係は、アテンションモデルを適用した場合でも問題なく、むしろ改善する方向であることが確認できた。その結果を基に、1 年目に基礎検討したビット線や読み出し回路の階層化のパ ターンを何種類か用意すれば、スパース化を活かした電源遮断が可能なことが確認できた。現在の方法を継続しデータを増やしていけば当初の計画通り、次元削 減することを実現するために必要な 1x1 畳込みを階層化で実現する CIM 設計が可能なことがわかった。

最終年度のまとめとしては、本研究の目的はモデルアーキテクチャの観点から無駄な計算を削減しハードウエア的に実装しやすい手法を探ることである。コンボリューショナルニューラルネットワークにおいてチャネル方向の次元に冗長性が大きく、この方向の次元削減が、１）精度を落とさずパラメータを削減できる最も効率の良い手段である、２）ハードウエア実装の点でも貢献度が少ないチャナネルを刈り取ることは比較的容易であるとの検討結果を得た。本研究ではアテンション手法に注力し、１）チャネルアテンション、２）空間アテンションによる効果の確認を実施した。実験においては、パラメータ数を削減するのに欠かせない１ビットへの量子化を前提に実施した。１ビットの量子化により学習曲線が不安定になり誤差が減少するエポック数が大きくなる課題があるので、本研究では、アテンション機構を改良し学習曲線を高速化することも実現した。実際には SE-ResNet Attention アルゴリズムを用いて実験した。実験結果の考察から Attention すべき特徴マップの決定プロセスにバイナリー法を用いることで、Attention すべきベクトルの決定が速くなり学習曲線を高速化することにつながることを明らかにした。ResNet の構成を独自にコンパクトにした ResNet14 構成の 3 ブロックで平均チャネル削減割合が約 50%になり、パラメータの削減可能なことを明確にした。SVHN ヤ CiFAR10 のデータセットにおいても同様の結果が得られた。さらに、説明可能な効果を明らかにするために　Eigen CAM を用いて本研究手法が空間的にも適切な Attention を獲得していることを確認した。この結果は査読付き 6 本の学術論文、査読付き 2 本の国際学会発表、そして、2 本の国内学会発表につながった。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　4件／うちオープンアクセス　2件）

１．著者名 Jiazheng Xi and Hiroyuki Yamauchi	４．巻 13
２．論文標題 Relaxed Training Procedure for a Binary Neural Network	５．発行年 2023年
３．雑誌名 International Journal of Machine Learning and Computing	６．最初と最後の頁 7-12
掲載論文のDOI（デジタルオブジェクト識別子） 10.18178/ijml.2023.13.1.1124	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Peng Yu and Hiroyuki Yamauchi	４．巻 ICECC 4
２．論文標題 A Machine Learning Based Fuel Consumption Saving Method with Time and Environment Dependency Aware Management	５．発行年 2022年
３．雑誌名 ACM digital library	６．最初と最後の頁 40-49
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3531028.3531035	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Jiazhen Xi and Hiroyuki Yamauchi	４．巻 35
２．論文標題 A layer-wise ensemble technique for binary neural network	５．発行年 2021年
３．雑誌名 International Journal of Pattern Recognition and Artificial Intelligence	６．最初と最後の頁 1-21
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S021800142152011X	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Peng Yu, Jiazheng Xi and Hiroyuki Yamauchi	４．巻 36
２．論文標題 Time and Environment Dependency Aware Fuel Consumption Tracking Method for Improving Drivers and Trucks Management	５．発行年 2021年
３．雑誌名 The 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2021)	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ITC-CSCC52171.2021.9501436	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Shaoqing Wu, Hiroyuki Yamauchi	4．巻 -
2．論文標題 A Speed-up Channel Attention Technique for Accelerating the Learning Curve of a Binarized Squeeze-and-Excitation (SE) based ResNet Model	5．発行年 2024年
3．雑誌名 Journal of Advances in Information Technology	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Shaoqing Wu, Hiroyuki Yamauchi	4．巻 -
2．論文標題 A Binarized Feature Mapping Technique for Enhancing Squeeze-and-Excitation (SE) Mechanism	5．発行年 2024年
3．雑誌名 International Journal of Machine Learning	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

1．発表者名 Zhufeng Li, Hiroyuki Yamauchi
2．発表標題 Image Recognition Accuracy, Number of Parameters and Computational Complexity Using Channel Reduction by Dimensional Compression and Attention Function
3．学会等名 6th International Conference on Electronics, Communications and Control Engineering (ICECC 2023)（国際学会）
4．発表年 2023年

1．発表者名 Li Zhufeng, Guan Weijie, Hiroyuki Yamauchi
2．発表標題 Graph Structure Exploration for Reinforcement Learning State Embedding -- Train Tetris Agent with Graph Neural Network
3．学会等名 6th International Conference on Electronics, Communications and Control Engineering (ICECC 2023)（国際学会）
4．発表年 2023年

１．発表者名 李　鋳峰，山内寛行
２．発表標題 次元圧縮とアテンション機能によるチャネル削減を利用した画像認識精度とパラ メータ数と計算量の削減手法の最適化の検討
３．学会等名 2022 年度第75回電気・情報関係学会九州支部連合大会
４．発表年 2022年

１．発表者名 呉 少卿，李 鋳峰，山内 寛
２．発表標題 SEモジュールによるチャネルアテンション効果のモデルEPOCH・チャネル幅・深さの依存性考察
３．学会等名 2023 年度第75回電気・情報関係学会九州支部連合大会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関