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　４ 年　 ５ 月 １０ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４０７９

研究課題名（和文）近似コンピューティングを活用した深層ニューラルネットワークアクセラレータの開発

研究課題名（英文）Development of Deep Neural Network Accelerator Utilizing Approximate Computing

研究代表者
劉　載勲（YU, JAEHOON）

東京工業大学・科学技術創成研究院・准教授

研究者番号：７０７２６９７６

交付決定額（研究期間全体）：（直接経費）　　13,500,000円

研究成果の概要（和文）：本研究では3年間の研究期間において、深層ニューラルネットワークの学習と推論における近似コンピューティング手法を考案し、それをサポートする演算回路と推論アクセラレータを提案した。またその成果を国際会議６件と論文誌2件にて公表している。特筆すべき成果としては、チップのオリンピックと呼ばれるISSCC2022に発表された推論アクセラレータHiddeniteがあげられる。乱数重みを用いることで深層ニューラルネットワークのメモリ要求を大幅に減らしたHiddeniteは40nmの比較的古いプロセスで実装されているにも関わらず、最先端プロセスの推論アクセラレータと同等以上の処理効率を示した。

研究成果の学術的意義や社会的意義
本研究成果の学術的意義は、 深層学習のアルゴリズムから、アーキテクチャ、回路技術、設計技術までをカバーしたクロスレイヤー型研究による解析と最適化を行い、深層ニューラルネットワークにおいて不必要な冗長性と厳密性を取り除くためにどのようなアプローチが有効であるかを明らかにしたことにある。
またこれにより深層ニューラルネットワークを利用するために必要な計算リソースと電力リソースの制約を緩和することが可能となり、それが適用可能な 範囲を大きく広げた点で大きな社会的意義を持つ。

研究成果の概要（英文）：We devised approximate computing methods for learning and inference of deep neural networks during the three-year research period. Also, we proposed an arithmetic circuit and an inference accelerator to support them. These results have been published in six international conferences and two journals. One of the most notable achievements is Hiddenite, an inference accelerator presented at ISSCC2022, the Olympics of Chips. Hiddenite significantly reduces the memory requirements of deep neural networks by using random weights. We implemented Hiddenite on a relatively old 40nm process. Yet, it showed processing efficiency equivalent to or better than inference accelerators using state-of-the-art processes.

研究分野：AIプロセッサ

キーワード：深層ニューラルネットワーク　近似コンピューティング　深層学習　推論アクセラレータ

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

(1) 深層学習

近年の深層学習では、AlexNet や ResNet などの DNN の登場により学習性能が飛躍的に向上した。一方で、深層学習を実問題に応用するにはニューラルネットワーク (NN) の学習と推論に要する膨大な計算量の問題を解決する必要がある。現在の深層学習の研究では、NVIDIA 社の GPU (graphics processing unit) や Google 社の TPU (tensor processing unit)などのハードウェアアクセラレータを計算に用いている。しかし、NN の複雑化に伴って指数爆発的に増加する計算量はハードウェアアクセラレーションによる高速化効率を大きく上回っており、研究の深化と実用化を阻害している。大規模なデータセンタを離れ、限られた計算資源を持つエッジ端末に注目すると、比較的小規模な AlexNet でも 480MB のメモリと画像 1 枚あたり 720MFLOPS の計算量を要するため、深層学習の活用は計算資源、計算量ならびに電力の観点から絶望的な状況にある。

DNN の計算量問題を解決すべく、深層学習で構築した NN の冗長性と非厳密性に注目して、ネットワーク情報と計算を圧縮・削減・近似する手法の研究が始まっている。例えば、ネットワーク各層での入力とカーネルのスパース性を利用するプルーニングやカーネルスパース性の均一化による CNN (convolutional neural network) 処理の効率化 [Mao、 CVPR2016]、ネットワーク情報を量子化する Deep Compression [Han、 ICLR2016]、動的固定小数点 (dynamic fixed point) [Courbariaux、 ICLR2015]などの研究によって、NN に求められる情報量、計算量、計算資源をアルゴリズムレベルで数十分の一までに削減できることが示されている。

しかし、上記の NN のスパース性や非厳密性の活用は、密行列演算を対象とする GPU や TPU が得意とする処理ではない。そこで、スパースな FC (fully-connected) レイヤーの処理を目的とする EIE (efficient inference engine) [Han、 ISCA2016]や SCNN (sparse convolutional neural network) 向けハードウェアアーキテクチャ [Mukkara、 ISCA2017] など DNN の一部に注目した研究が始まっており、既存の GPU や TPU に比べて大幅な処理効率の改善が期待されている。しかし、3 次元テンソル間の畳み込み演算を 1 次元ベクトル間の積和演算に置き換えることを前提にしているため、大量のオフチップメモリアクセスを要求し、組込みシステムの電力上限を満足できない。小規模な手書き文字認識しかできないハードウェアや認識精度の低下を度外視した提案はいくつかあるが [例: Nurvitadhi、 ICFPT2016; Rastegari、 ECCV2016]、実用的に意味のある DNN への適用は極めて難しい。

(2) 近似コンピューティング

一方で計算品質と計算資源 (計算時間、消費電力など) のトレードオフを積極的に活用し、限られた計算資源で通常では達成できなかった計算量を実現する近似コンピューティングの研究が盛んに行われている。例えば、[Chippa、 TVLSI2014] の k-means クラスタリングの例では 5%の分別精度の低下で計算に必要なエネルギーを 1/50 に削減している。[Grigorian、 HPCA2015] は、NN の活用により 5%以下の誤差で GPU に対して 26 倍の加速を実現している。

近似コンピューティングの研究はソフトウェアからハードウェアまで広範囲に広がっているが、近似が許容できるアプリケーションドメインが限られる、アプリケーション毎に最適な戦略が異なる、オーバーヘッドが大きくスケーラビリティがない、計算品質と電力のトレードオフの質が低い、などの重要課題が残されている。DNN と近似コンピューティングの相性は良いと考えられるが、特に設計最適化されることもなく組み合わせてみたという事例が報告されているのみである [Wang、 TVLSI2017]。

２．研究の目的

本研究は GPU よりも 3 桁高いエネルギー効率を達成する DNN ハードウェアアクセラレータを開発することを目的とする。アルゴリズム、アーキテクチャ、回路技術、設計技術を跨いだクロスレイヤー最適化で、データ移動を最小化し、ネットワーク冗長性と計算厳密性を極限まで取り除くことで、計算エネルギー効率を飛躍的に高める。

３．研究の方法

(1) 近似コンピューティングアルゴリズムの開発

R1 年度は、エッジデバイスに適した軽量な機械学習手法として、単一パーセプトロンであるサポートベクタマシン（SVM）に対し、深層 NN を教師モデルとする知識の蒸留を行い、軽量な機械学習手法の精度向上を試み、その可能性を確かめた。

また、深層 NN の学習時に必要な訓練データの要求を軽減するために、GAN による訓練データの強化を行い、同程度の推論精度を得るために必要な訓練データの数を大幅に削減、または、同程度の訓練データ数で同等以上の推論精度を達成できることを明らかにした。

R2 年度は、近似コンピューティングに基づき、計算コストと推論精度を調整可能な単一 NN の表現に関する研究を行った。ProgressiveNN と命名した提案手法では、NN を構成する各重みの２進数表現における {0、1} を、{-1、+1} の対称的な量子化表現として扱うことで、単一の重み

から任意の N ビット近似表現を追加コストなしで利用できる特徴を持つ。R2 年度の研究では、対称的量子化表現とバッチ正規化係数のみの再学習によって、単一 NN から複数計算精度の NN を追加コストなしで利用できることを明らかにした。

(2) 近似コンピューティングに基づく低電力乗算回路の開発
深層 NN において支配的な消費電力を占める乗算処理に着目し、それにかかる消費電力を大幅に削減できる対数近似に基づく乗算回路の研究を行った。具体的には、 浮動小数点の乗算における仮数部の乗算を加算に置換し、それを用いて学習した NN の推論精度とそのときの電力効率を明らかにした。

(3) VLSI 実装
深層 NN 処理におけるエネルギー効率の向上を目指し、近似コンピューティングに基づく推論アクセラレータのアーキテクチャを明らかにし、それを 40nm プロセスで試作した。試作した VLSI は数 100mW で ResNet などの NN が高速動作できることを明らかにした。

４．研究の成果

(1) R1 年度
深層学習のアルゴリズムおよび実装における基礎的な研究を行い、その成果を国際会議３件の発表を行った。

アルゴリズムの研究では、深層 NN で用いられる蒸留の概念をサポートベクタマシン(SVM)やアンサンブル学習などの他機械学習に導入する試みを行い、 精度向上の可能性を確かめた。その結果、NN を教師モデルとする SVM への蒸留で約 2.8%の精度改善を確認し、APSIPA にて発表している。これは局所的ではあるが、NN の作る空間情報を他機械学習の実装形態を借りて近似できる可能性を示したものである。

次に深層 NN の学習における計算量削減の試みとして精度劣化を伴わない訓練データの削減手法について検討を行った。既存のデータクリーニングでは、不要データを削減するデータ削減前に学習されたモデルとデータ削減後に学習されたモデル間の距離を比較することによってデータの必要性を判断する。本研究では、サポートベクタマシンの学習時に選ばれるサポートベクタを重要度の高いデータとしてデータ削減を行うことを検討した。その結果、比較的単純なデータではその効果を一部確認することができたが、実用的な手法の発見には至っていない。本成果は上記と同様に APSIPA にて発表を行った。

最後に実装面での研究では、NN で必要な積和演算の近似手法として浮動小数点の対数近似乗算器を用いた実装方式を提案した。NN の学習では、推論に比べて広い範囲の値を扱う必要がある。浮動小数点の対数近似乗算はその要求を満たす近似計算であり、単純なデータセットを用いた学習では最大 1.54 倍の電力効率を達成した（図 1）。本成果は PATMOS にて発表を行った。

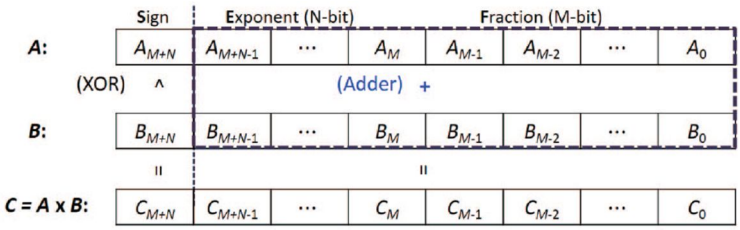

図 1　対数近似に基づく乗算処理
（Logarithmic approximate multiplier; LAM）

(2) R2 年度
深層 NN における精度と計算量を選択的に調整できる学習・推論アルゴリズムと、近似計算に基づくハードウェア実装方式に関する研究を行い、その成果を国際会議２件と論文誌１件として発表した。

まず学習・推論アルゴリズムの研究では、ProgressiveNN と名付けた NN の新たな学習・推論手法を提案して一度の学習で得られるネットワークモデルから複数計算精度の推論処理を実現し、少ないメモリオーバーヘッドで NN の推論精度と計算量を選択可能にした（図 2）。実験結果ではこれにより、8 ビット重みで学習された単一の NN を、1 ビットから８ビットまでの計算精度を 1 ビット単位で調整でき、CIFAR-10/100 における 8 ビットと 1 ビットでの推論精度差をそれぞれ 4%、10%未満に抑えることに成功した。本成果は国際会議（CANDAR）にて発表を行い、その改良とハードウ

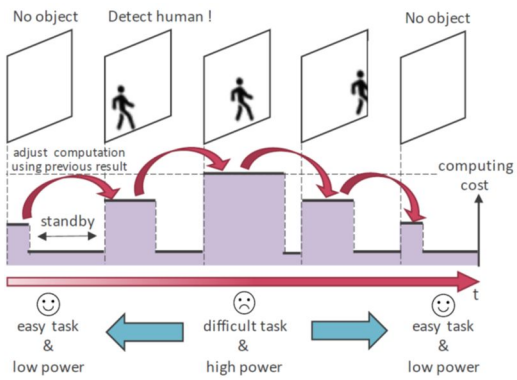

図 2　ProgressiveNN を用いたタスクに応じた動的な計算量制御の概要

ェア（HW）実装に関する内容を含め、論文誌（IJNC）に投稿し、条件付き採録の結果を得ている。
　次に R1 年度に提案した対数近似に基づく浮動小数点乗算器を改良し、その研究成果は論文誌（VLSI Journal）にて発表を行った。また適応的ビット幅と電圧スケーリングに基づく深層学習のエネルギー最小化手法を提案し、その成果を国際会議（ISCAS）にて発表している。

(3) R3 年度

R3 年度には、ProgressiveNN におけるビットシリアル演算に特化した積和演算回路の検討を行い、ProgressiveNN とその演算回路をまとめて、論文誌（IJNC）1 件として発表している。

　また、重みを乱数初期化状態からアップデートでせず、スーパーマスクによって機能する部分ネットワークを選び出すことによって学習を行う隠れ NN [Ramanujan、 CVPR2020]をベースとし、それに適したアーキテクチャを提案し、40nm プロセスの VLSI として実装した。Hiddenite と命名した推論アクセラレータは、40nm プロセスでありながら他の先進的なプロセスを用いた推論アクセラレータと同等以上の性能を達成している。本成果は正解的にも高く評価され、 一線級の国際会議であり、チップ分野のオリンピックと呼ばれる ISSCC2022 に採択・発表されている。

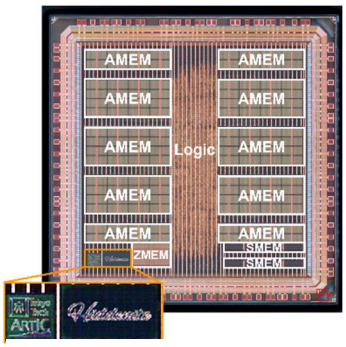

図 3　推論アクセラレータ Hiddenite

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Cheng TaiYu、Masuda Yukata、Chen Jun、Yu Jaehoon、Hashimoto Masanori	４．巻 74
２．論文標題 Logarithm-approximate floating-point multiplier is applicable to power-efficient neural network training	５．発行年 2020年
３．雑誌名 Integration	６．最初と最後の頁 19～31
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.vlsi.2020.05.002	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 Suzuki Junnosuke、Kaneko Tomohiro、Ando Kota、Hirose Kazutoshi、Kawamura Kazushi、Chu Thiem Van、Motomura Masato、Yu Jaehoon	４．巻 11
２．論文標題 ProgressiveNN: Achieving Computational Scalability with Dynamic Bit-Precision Adjustment by MSB-first Accumulative Computation	５．発行年 2021年
３．雑誌名 International Journal of Networking and Computing	６．最初と最後の頁 338～353
掲載論文のDOI（デジタルオブジェクト識別子） 10.15803/ijnc.11.2_338	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6件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 TaiYu Cheng, Jaehoon Yu, Masanori Hashimoto
２．発表標題 Minimizing Power for Neural Network Training with Logarithm-Approximate Floating-Point Multiplier
３．学会等名 IEEE International Symposium on Circuits and Systems（国際学会）
４．発表年 2020年

１．発表者名 Junnosuke Suzuki ， Kota Ando ， Kazutoshi Hirose ， Kazushi Kawamura ， Thiem Van Chu ， Masato Motomura ， Jaehoon Yu
２．発表標題 ProgressiveNN: Achieving Computational Scalability without Network Alteration by MSB-first Accumulative Computation
３．学会等名 International Symposium on Computing and Networking (CANDAR)（国際学会）
４．発表年 2020年

１．発表者名
TaiYu Cheng ; Jaehoon Yu ; Masanori Hashimoto

２．発表標題
Minimizing Power for Neural Network Training with Logarithm-Approximate Floating-Point Multiplier

３．学会等名
2019 29th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)（国際学会）

４．発表年
2019年

１．発表者名
Shota Fukui ; Jaehoon Yu ; Masanori Hashimoto

２．発表標題
Distilling Knowledge for Non-Neural Networks

３．学会等名
2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2019年

１．発表者名
Toranosuke Tanio ; Kouya Takeda ; Jeahoon Yu ; Masanori Hashimoto

２．発表標題
Training Data Reduction using Support Vectors for Neural Networks

３．学会等名
2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2019年

１．発表者名
Hirose Kazutoshi、Yu Jaehoon、Ando Kota、Okoshi Yasuyuki、Garcia-Arias Angel Lopez、Suzuki Junnosuke、Chu Thiem Van、Kawamura Kazushi、Motomura Masato

２．発表標題
Hiddenite: 4K-PE Hidden Network Inference 4D-Tensor Engine Exploiting On-Chip Model Construction Achieving 34.8-to-16.0TOPS/W for CIFAR-100 and ImageNet

３．学会等名
2022 IEEE International Solid- State Circuits Conference (ISSCC)（国際学会）

４．発表年
2022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
研究分担者	橋本　昌宜 (Hashimoto Masanori) (80335207)	京都大学・情報学研究科・教授 (14301)	

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