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

平成　２９　年　　６　月　　４　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３００２７

研究課題名（和文）画像認識向けニューラルネットワークプロセッサの研究

研究課題名（英文）Study on Neural Network Processor for Image Recognition

研究代表者
廣本　正之（HIROMOTO, Masayuki）

京都大学・情報学研究科・助教

研究者番号：６０７１８０３９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究ではニューラルネットワークを用いた画像認識向けプロセッサの高効率化に向け，認識処理内部の演算の冗長性に着目し，不正確な演算を積極的に活用する計算手法を提案した．近似演算の利用や低電圧動作等を行うことでハードウェアのエネルギー効率が向上する一方，演算誤差により認識性能に影響を与える可能性がある．本研究ではこれらの演算誤差を許容する学習アルゴリズムを考案し，認識精度を維持したままハードウェアの性能電力比の向上を実現した．

研究成果の概要（英文）：This work proposes efficient computing methods for image recognition by neural network processors, in which inaccurate computation is positively utilized based on the observation of the calculation redundancy in the recognition process. For the inaccurate computation, approximate computing and low-voltage operation of the circuits can be used to improve their energy efficiency. However, they may affect the recognition performance because of the calculation errors. In this work, novel algorithms to mitigate the calculation errors are proposed, and they achieved the energy efficiency improvement of the neural network processor while keeping the recognition accuracy.

研究分野：　計算機システム

キーワード：　画像認識　ニューラルネットワーク　ディープラーニング　低消費電力設計　近似計算　メモリスタ

１．研究開始当初の背景

実世界の物事をコンピュータに理解させる認識技術は，今や様々な分野において欠かせない技術である．特に画像認識技術は，監視，ロボティクス，車載等の様々な分野において今後益々重要になり発展が期待される技術である．近年は統計的手法や機械学習を用いた認識アルゴリズムが多数提案されているが，その中でとりわけ注目が集まっているのがディープラーニングと呼ばれる機械学習手法である．ディープラーニングは，脳の神経回路の構造を模擬した数理モデルであるニューラルネッーワークの一種であり，より複雑で深い構造を用いることで高精度な認識を可能とする手法である．

しかし，ディープラーニングは高精度な認識を実現できる反面，従来の認識アルゴリズムに比べて多大な計算量を必要とする．これは認識システムの消費電力の増大につながり，特に電力制約の厳しい組込み機器等においては大きな課題となる．実用的な認識システムを構築するためには，アルゴリズムの改善だけでなく，そのソフトウェア，ハードウェア実装に関しても複合的に検討する必要がある．

２．研究の目的

本研究の目的は，ニューラルネットワークを用いる画像認識処理に適したプロセッサアーキテクチャ及び画像認識システムを明らかにすることである．本研究では上記のディープラーニングに着目し，ニューラルネットワークを用いた画像認識向けプロセッサについて，アルゴリズムとハードウェアの両面からの検討を行う．具体的には，(1) ディープラーニングの利用による高精度な認識処理，(2) あいまいなコンピューティングの利用による高い性能電力効率，(3) 半導体プロセスのばらつきや経年変化を補償するコンピューティングの 3 つを特徴とする，性能電力性能の優れた画像認識システムの実現を狙う．

３．研究の方法

上記目的を達成するため，以下の 3 つの方法により研究を実施した．

(1) ディープラーニングの利用による高精度な認識処理：深い階層構造を持つニューラルネットワークを用いることにより，高精度な画像認識処理を実現することができる．まずはディープラーニングのアルゴリズムを分析し，画像認識性能およびそれを実現するために必要な計算量を評価する．また，それに適した並列計算機構を備えたプロセッサアーキテクチャを検討する．

(2) あいまいなコンピューティングの利用による高い性能電力効率：画像等の認識処理は，途中の計算過程が多少不正確であっても最終的な結果にそれほど影響を与えない．従って，積極的に不正確な（あいまいな）計算を取り入れることにより，認識精度を保ったまま計算効率を向上させられる可能性がある．本研究ではニューラルネットワークの特性を活かしたあいまい計算手法を検討し，上記プロセッサに適用することで性能電力効率の向上を狙う．

(3) 半導体プロセスのばらつきや経年変化を補償するコンピューティング：回路を構成する個々のトランジスタは，製造時のばらつきや経年変化により不均一な特性を持つ．トランジスタを定格電圧以下の低電圧で動作させるとエネルギー効率が向上することが知られているが，上記のばらつきにより回路を正常動作させることが困難である．本研究ではあいまい計算を活用することにより，低電圧動作時の不正確な動作を許容する手法を検討し，認識システムのさらなる効率向上を狙う．

４．研究成果

(1) ニューラルネットワーク向けハードウェアプラットフォームの構築

ニューラルネットワークを利用した画像認識を高効率に実行するプロセッサを実現するため，その基本となるハードウェアプラットフォームの構築を行った．プラットフォームの概要を図 1 に示す．

提案プラットフォームはソフトウェアフローとハードウェアフローの 2 つから構成されている．ソフトウェアフローでは各種ニューラルネットワークのアルゴリズム自体の認識性能評価を行うことができ，ハードウェアフローではそれを実現するハードウェアの性能を速度，面積，消費電力の観点で評価することができる．本プラットフォームにより，ニューラルネットワークを用いる一般的な画像認識アルゴリズムの認識精度や処理性能等を，ソフトウェアとハードウェアの両面から定量的に評価することが可能となる．

また本プラットフォームでは，画像認識においてよく利用される畳み込みニューラルネットワーク(CNN)を対象としており，それを効率良く実行できるハードウェアをベースアーキテクチャ(図中の Base Arch.)として採用している．ベースアーキテクチャの概要を図 2 に示す．CNN における主要な処理である畳み込み(Conv)，バイアス加算(Bias)，活性化関数(ReLU)，プーリング(Pool)といった要素が，それぞれ対応するモジュールとして相互接続される形で構成されている．この構成により，様々なネットワークに柔軟に対応でき，また次節で述べる高効率化手法を容易に適用，評価することが可能となる．

本研究成果は査読付き国際会議等にて論文発表を行っている(①②)．CNN 向けのハー

ドウェアはここ数年間，国内外において盛んに研究されているが，本研究成果はそれらと遜色の無い性能を示している．また，多様なネットワークに柔軟に対応できる点は大きな特徴であり，他に類を見ないものである．今後の展望としては，本プラットフォームを活用することにより，次節で述べる近似演算手法を含め，より効率の良いソフト・ハード協調設計を実現することが期待される．

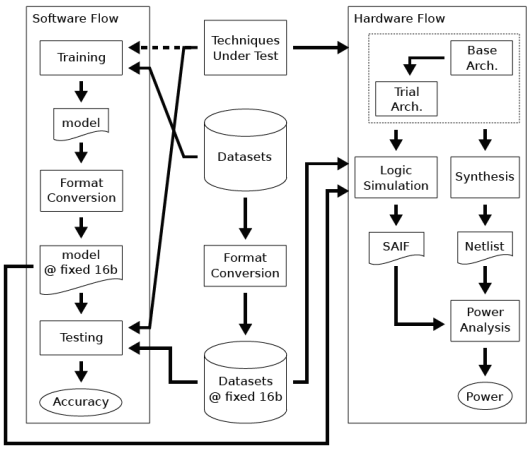

図 1　提案ハードウェアプラットフォーム

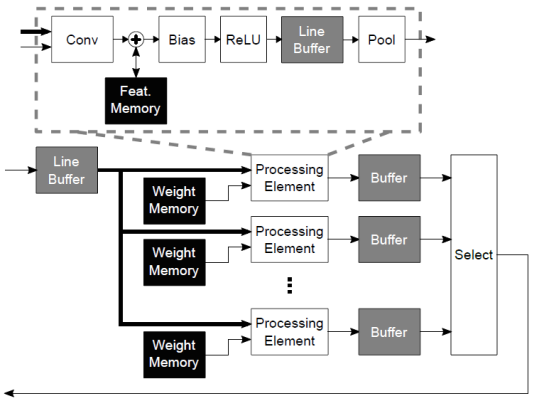

図 2　CNN 向けハードウェアアーキテクチャ

(2) 近似演算の活用によるニューラルネットワークハードウェアの高効率化

ニューラルネットワークの内部演算には一定の冗長性があり，多少あいまいな計算を行っても最終的な認識精度には大きな影響を与えない可能性がある．本研究ではハードウェア化に適した近似演算を CNN に適用することで，CNN の畳み込み演算にかかるコストを低減させる手法を提案した．

提案手法では畳み込み演算とプーリングの組合せに着目し，出力への影響の少ない中間演算に対し近似および省略を行う．CNN においては畳み込み演算とプーリングは順番に行われるが，図 3 のようにプーリングでは複数の畳み込み演算結果のうち，最終的に 1 つの結果のみが出力される．従ってプーリングで選択されない畳み込み演算結果は，出力に直接の影響を与えないことになる．そこで本研究では図 4 のように，まず演算コストの小さい近似演算(AppConv)によりプーリングの出力を予測し，予測結果に基づいて必要な畳み込み演算のみを実行する二段階の計算手法を提案した．これにより必要な畳み込み演算数をプーリング領域のサイズに応じて(図中の例では 1/4 に)低減できた．提案手法のソフトウェア・ハードウェア実装を行った結果，認識精度をほとんど損なうことなく CNN ハードウェアの消費電力を約 20%低減できることを示した(図 5)．

本研究成果は査読付き国際会議等にて論文発表を行っている(⑤⑥)．CNN に対する近似演算手法についてもここ数年で国内外での研究が活発化しているが，それらは主にアルゴリズム面での工夫が主である．本研究での提案手法は実際にハードウェア化した際に有効な手法であり，本成果は特に組込みシステムへの応用において有用であると考える．今後の展望としては，プーリング以外の処理に対しても同様の予測・近似手法を適用することにより，さらなる高効率化を図ることが期待される．

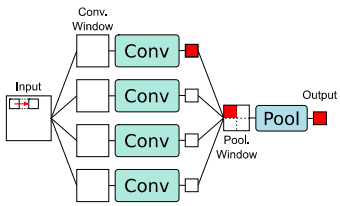

図 3　CNN の畳み込み演算とプーリング

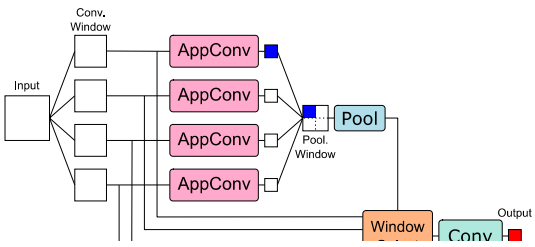

図 4　提案手法の畳み込み演算とプーリング

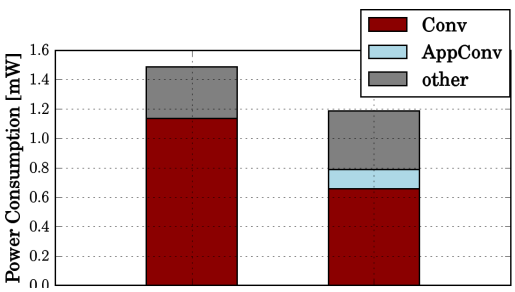

図 5　提案手法による電力削減効果

(3) 低電圧動作時の演算誤りを緩和するニューラルネットワークアルゴリズム

上述の近似を用いたものとは異なるアプローチの高効率化手法として，演算回路の超低電圧動作が挙げられる．回路が正常に動作する定格電圧よりさらに電圧を下げることにより，回路動作時のエネルギー効率が向上することが知られている．しかし過度な低電圧化は各トランジスタの遅延時間を増大させ，演算回路の出力に誤りを発生させることになる．従って，このような回路をそのままニューラルネットワークハードウェアに適用すると，発生する演算誤りの程度によっては正しく認識処理が行えない可能性がある．

本研究では，このようなデバイスの物理特性に起因して発生する演算器の誤りを，ニューラルネットワークのアルゴリズムの工夫により緩和し，定格電圧以下でのニューラルネットワークハードウェアの動作を可能とする手法を提案した．提案手法では前節と同様に CNN を対象とし，CNN 中に含まれるプーリング処理に中央値演算を採用した．一般的なプーリング処理では最大値や平均値が用いられることが多いが，低電圧動作により一部の値に大きな演算誤差が発生した場合，それらの最大値や平均値も大きく影響を受けることになる．これに対し，提案手法では中央値を採用することで，大きな演算誤差を含む値の影響を小さくすることが可能になる．中央値プーリングを採用することで，画像認識の精度を損なうことなく，畳み込み演算器の動作電圧を約 10%低減できることをシミュレーションにより確認した(図 6)．

本研究成果は査読付き国内会議等にて論文発表を行っている(⑦⑧)．回路の低電圧動作についての研究は以前から国内外で盛んに行われているが，本研究はニューラルネットワークのアルゴリズムに着目した新たなアプローチである．今後の展望としては，低電圧動作時の誤り特性をモデル化し，ニューラルネットワークの学習時にその特性を考慮することで，さらなる低電圧動作を実現することが期待される．

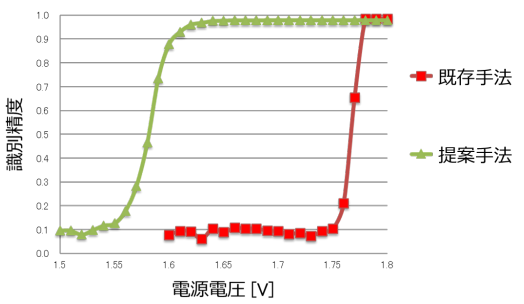

図 6　動作電圧と認識精度の関係

(4) ばらつきを考慮したメモリスタニューラルネットワークの学習

前節までの検討内容はディジタル回路を用いた高効率化に関するものであったが，アナログ回路を用いることで消費電力を大幅に削減できる可能性がある．本研究では特に，メモリスタと呼ばれる素子を用いたニューラルネットワークハードウェアについて検討を行った．メモリスタは，通過した電荷量に応じその抵抗値が変化する特性を持つ受動素子であり，メモリと演算器の機能を兼ねさせることができる．図 7 のようにメモリスタを格子状に接続したアレイ回路を用いることにより，入力電圧 V とメモリスタのコンダクタンス行列 G との積和演算結果を，電流値 I として取り出すことが可能となる．この計算は微小な電流により実現できるため，ニューラルネットワーク中の積和演算を極めて低い消費電力で行うことが可能となる．しかし実際の各素子のばらつきは非常に大きく，所望の学習を行うことが難しいという課題がある．

そこで本研究ではメモリスタのばらつきを考慮したシミュレーション環境を構築し，素子のばらつきがニューラルネットワークの学習に与える影響を評価した．その結果，図 8 に示すように，5 通りのばらつきにおいてそれぞれ異なる収束性を示すことが分かり，素子特性に応じた学習パラメータの調整が必要であることが明らかになった．

本研究成果は国内研究会にて論文発表を行っている(④)．メモリスタを用いたニューラルネットワークに関する研究はここ数年で活発になりつつあるが，素子のばらつきまでを考慮した事例は未だ少なく，本研究の寄与する所は大きいと考えられる．今後の展望としては，さらに大規模なニューラルネットワークを対象としたシミュレーション技術を確立し，メモリスタニューラルネットワークの実現性を示すことで新たなコンピューティング手法を実現することが期待される．

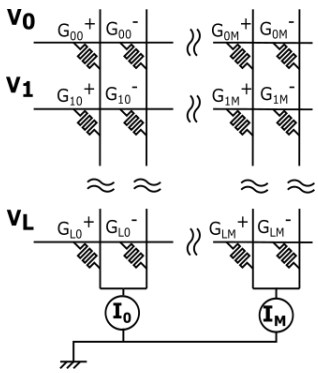

図 7　メモリスタアレイによる積和演算回路

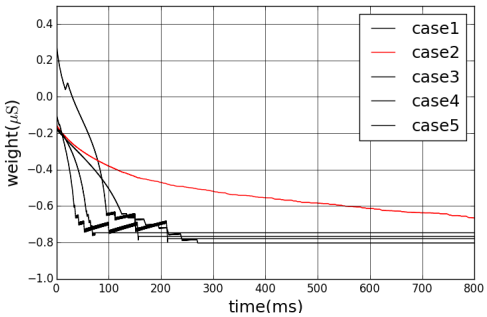

図 8　メモリスタニューラルネットの学習収束性

5．主な発表論文等

〔学会発表〕（計 9 件）

① 氏家 隆之，廣本 正之，佐藤 高史：演算簡略化手法評価のための畳み込みニューラルネットワークの FPGA 実装，第 42 回パルテノン研究会，pp.51-56，2016 年 12 月 17 日，東海大学高輪キャンパス（東京都港区）．

② Takayuki Ujiie, Masayuki Hiromoto, and Takashi Sato: Hardware Accelerator of Convolutional Neural Network for Image Recognition and its Performance Evaluation Platform, the 20th Workshop on Synthesis And System Integration of Mixed Information technologies (SASIMI2016), pp.16-17, 24 Oct. 2016. 京都リサーチパーク（京都府京都市）．

③ 三宅 哲史，氏家 隆之，廣本 正之，佐藤 高史：Binarized Neural Network を用いた画像認識ハードウェアの消費エネルギー評価，電子情報通信学会ソサイエティ大会，p.64，2016 年 9 月 20 日，北海道大学（北海道札幌市）．

④ 山森 聡，廣本 正之，佐藤 高史：ばらつきを考慮したメモリスタモデルによるニューラルネットワークの学習収束性の評価，電子情報通信学会ソサイエティ大会，p.85，2016 年 9 月 23 日，北海道大学（北海道札幌市）．

⑤ Takayuki Ujiie, Masayuki Hiromoto, and Takashi Sato: Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.870-876, 1 July 2016, Las Vegas (USA).

⑥ 氏家 隆之，廣本 正之，佐藤 高史：近似的予測戦略に基づく畳み込みニューラルネットワークプロセッサの低電力化，第 29 回 回路とシステムワークショップ，pp.13-18，2016 年 5 月 12 日，北九州国際会議場（福岡県北九州市）．

⑦ 氏家 隆之，大荷 唯明，廣本 正之，佐藤 高史：低電圧畳み込みニューラルネットワーク回路における演算誤り緩和に向けたプーリング手法の検討，電子情報通信学会ソサイエティ大会，A-3-8，p.53，2015 年 9 月 10 日，東北大学川内キャンパス（宮城県仙台市）．

⑧ 大荷 唯明，廣本 正之，佐藤 高史：ニューラルネットワークハードウェアの低電圧動作時における演算誤り緩和，第 28 回 回路とシステムワークショップ，pp.249-254，2015 年 8 月 4 日，淡路夢舞台国際会議場（兵庫県淡路市）．

⑨ Masayuki Hiromoto and Takashi Sato: A Case Study of Chinese Calligraphic Style Classification Using Deep Neural Network, International Workshop on Smart Info-Media Systems in Asia (SISA), 10 Oct. 2014, Ho Chi Minh City (Vietnam).

6．研究組織

(1)研究代表者

廣本　正之　（HIROMOTO, Masayuki）
京都大学・大学院情報学研究科・助教
研究者番号：60718039

(2)研究分担者

廣本　正之　（HIROMOTO, Masayuki）
京都大学・大学院情報学研究科・助教
研究者番号：60718039

(3)連携研究者

佐藤　高史　（SATO, Takashi）
京都大学・大学院情報学研究科・教授
研究者番号：20431992