科学研究費助成事業（基盤研究（Ｓ））公表用資料
〔令和２(2020)年度 中間評価用〕

平成３０年度採択分
令和２年３月３１日現在

知能コンピューティングを加速する自己学習型・革新的アーキテクチャ基盤技術の創出

Innovative Self-Learnable Architecture Platform for Accelerating Intelligent Computing

課題番号：18H05288

本村　真人（MOTOMURA, MASATO）

東京工業大学・科学技術創成研究院 AI コンピューティング研究ユニット・教授

研究の概要（４行以内）
DNN 処理エンジン，アニーリング計算機，ニューロモルフィック HW など広義の AI コンピューティングの研究成果を総合的に結集し、今後の知能コンピューティング応用分野に向けた計算基盤として，既存プログラマブル HW(=FPGA)を凌駕・置換する自己学習型・機能獲得型リコンフィギュラブル HW 基盤技術を提案する．

研　究　分　野：情報学

キ　ー　ワ　ー　ド：AI コンピューティング，深層ニューラルネット，ニューロモルフィック

１．研究開始当初の背景

深層ニューラルネットワーク(Deep Neural Network: DNN)の勃興により、AI(人工知能)技術とその社会応用が大きく進展している。AI 技術をより賢く進化させ、より低エネルギーで実現し、将来の超スマート社会を支える「知能コンピューティング」へと発展させていくためには、ソフトウェア技術だけではなく基盤となるハードウェア(HW)技術やアーキテクチャ技術の大きな進化が欠かせない。

２．研究の目的

本研究では、DNN 処理を加速する HW エンジンのアーキテクチャ技術を中核として、DNN の隣接領域であり、最適化問題を高速に解くアニーリング HW 分野やより脳に近い情報処理を目指したニューロモルフィック HW 分野の最新の知見や研究進展を積極的に結集して、将来の知能コンピューティングを支える革新的アーキテクチャ基盤技術の創出を目指す(図 1)。

３．研究の方法

DNN，アニーリング，ニューロモルフィックの三つの分野を対象とした上で，(a)知能コンピューティングを意識したリコンフィギュラブル HW アーキテクチャ，(b)アルゴリズム-アーキテクチャ協創による高効率な DNN 推論/学習方式，(c)アルゴリズム-回路の協創による高エネルギー効率 HW 方式、の三つの技術ベクトルで研究開発を推進する．研究代表者と 3 名の研究分担者が有するそれぞれ異なる強みを活かしながら，上記基盤技術の構築を進める。

４．これまでの成果

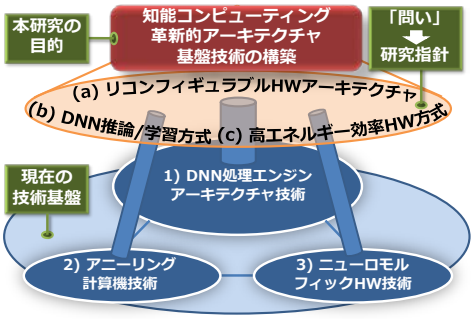

図 1. 研究構想の全体像

研究開始後，４名中３名の所属研究組織異動があったが，当初の緊密な協力関係を維持して研究を進めている．以下，図１の三分野毎にこれまでの成果の概略を記載する．

1)　DNN 関連技術

集積回路分野の最高峰国際会議である ISSCC2018 で発表した対数量子化・SRAM3 次元積層チップ(QUEST)に関して，アーキテクチャ詳細の開示とともに，様々な DNN ネットワーク形態の 2 次元演算器アレイへのマッピング手法や，量子化ビット数を変えたときの予測精度の振舞い，メモリ読み出し時間を変えたときの動作速度への影響などを総合的に評価したジャーナル論文を同最高峰ジャーナルである JSSC で発表した[1]．この業績は，独創的な 2 次元アレイ演算方式と対数量子化の実用性を明らかにした点で高く評価されている．

また，DNN 推論を加速する新たな手段として，出力活性化値がゼロになる無効なニュ

ーロンを予測して，無駄な DNN 演算を削減するコンセプトを提案し[2]，その有用性を評価した[3]．この提案では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．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が可能である．このアイデアは高く評価され，[2][3]それぞれ国内研究会で優秀講演として表彰された(国際会議投稿準備中)．

更に，画像処理分野における豊富な実績と知見を活かして，DNN の画像処理応用への展開も進めている．特に，一画素単位で輝度調整とノイズ抑制を実現する高効率画像処理ハードウェアを実現し，IEEE Transaction で発表した[4]．DNN 入力の前段に用いることで，より精度の高い画像認識が可能になる．

2) アニーリング計算関連技術

アニーリング計算技術は，スピンというバイナリ変数の結合体を用いて近似的に最適化問題を解く技術である．HW 実装の都合上，スピン間結合を局所的・規則的な形に制限したモデルが良く用いられるが，問題グラフのマッピングが難しいという問題がある．この問題の解決に向けて，問題グラフの時分割処理(すなわち HW のリコンフィギュレーション)によりマッピングの自由度を増やすことを提案し FPGA 上で実証した[5]．

また，2017 年に発表したバイナリ量子化 DNN アクセラレータのアーキテクチャを活かすことで，アニーリング計算を並列化できるという着想を得て，本課題内で研究を開始した．その後，数理物理学の専門家の協力も必要なことから，JST CREST にスピンアウトした別プロジェクトとしてその研究を進め，ISSCC2020 で全結合・全並列型のアニーリングプロセッサチップを発表した(参考情報)．

3) ニューロモルフィック関連技術

極めて高いエネルギー効率を実現する脳の情報処理からどう学びどう DNN 等に生かしていくかがここでの重要な命題である．

注力課題として，リザーバコンピューティングを取り上げ，原子スイッチを使ったリザーバで RNN(Recurrent NN)の置き換えを狙う研究を進めている．また，時系列信号間の相関により超小型の回路構成で近似計算が行える確率的コンピューティング手法に着目し，これを基本演算として DNN の Forward/Backward 計算を実現していく研究に着手した．まずは単純パーセプトロンを対象にフレームワーク構築を進めている．

５．今後の計画

研究対象としている三分野間で，テーマや技術の融合を進めていく．例えば，バイナリ DNN 研究で得た着想から生まれたアニーリング計算ハードウェア(ISSCC2020)のアイデアを，今度は DNN の学習やアーキテクチャ探索に展開する活動を進める．また，画像処理のドメイン知識を DNN に活かすことで，画像センサから画像認識までを一気通貫に最適化する研究に取り組む．更に，アニーリング計算とリザーバ計算の相互乗り入れも注目すべき研究課題と考えている．

DNN の分野の進展は目覚ましく，与えられた構成を HW 化するというスタンスでは研究として成立しない．最先端の学習・推論に関するアルゴリズム・アーキテクチャ・効率化手法等の技術の先を予想して計算基盤の構築を進める必要がある．その際，アニーリングやニューロモルフィックなど，少し先回りした基礎的な技術分野の知見を積極的にキャストすることが，真に基盤的で長期的な技術構築に重要である．研究体制内個々の強みを武器にして，今後も発見的な研究に取り組んでいく．

６．これまでの発表論文等（受賞等も含む）

[1] Kodai Ueyoshi, Kota Ando, Kazutoshi Hirose, Shinya Takamaeda-Yamazaki, Mototsugu Hamada, Tadahiro Kuroda, and Masato Motomura: QUEST: Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96-MB 3D SRAM Using Inductive Coupling Technology in 40-nm CMOS, IEEE Journal of Solid-State Circuits, Vol.54, pp.186-196, January 2019.

[2] "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，電子情報通信学会研究会報告 RECONF2019-18, pp.97-102, 2019 年 5 月 10 日発表．リコンフィギャラブルシステム研究会優秀講演賞 受賞．

[3] 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，電子情報通信学会研究会報告 CPSY2019-6, pp.51-56, 2019 年 6 月 11 日発表．情報処理学会システム・アーキテクチャ研究会若手奨励賞 受賞．

[4] Ambalathankandy, P., Ikebe, M., Yoshida, T., Shimada, T., Takamaeda, S., Motomura, M., and Asai, T. An adaptive global and local tone mapping algorithm implemented on FPGA. IEEE Transactions on Circuits and Systems for Video Technology, 10.1109, TCSVT.2019.2931510, 2019.

[5] Kasho Yamamoto, Masayuki Ikebe, Tetsuya Asai, Masato Motomura, and Shinya Takamaeda-Yamazaki: FPGA-based annealing processor with time-division multiplexing, IEICE Transactions on Information and Systems, Vol.E102-D, 2019.

７．ホームページ等

http://www.artic.iir.titech.ac.jp/