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　 ６ 月 １３ 日現在

機関番号：１４６０３
研究種目：挑戦的研究（萌芽）
研究期間：2022～2023
課題番号：２２Ｋ１９７７５

研究課題名（和文）主記憶転置型・超小型離散行列シリアル演算機構

研究課題名（英文）Ultra-small sparse matrix serial computation mechanism with memory transpose

研究代表者
中島　康彦（NAKASHIMA, YASUHIKO）

奈良先端科学技術大学院大学・先端科学技術研究科・教授

研究者番号：００３１４１７０

交付決定額（研究期間全体）：（直接経費）　　4,400,000円

研究成果の概要（和文）：第１に、新しいバイセクション・ニューラルネットワーク(BNN)に基づき、論理PEアレイが、物理計算ユニットに柔軟に分割写像される仕組みを考案した。論理PEアレイの形状や位置の調整により、機能に関する空間的再構成能力を実証した。精度に関する時間的再構成能力は、SCビットストリームの長さ調整により実現できた。エネルギー効率が他の最先端近似計算ユニットより優れていることを示した。第２に、確率的コンピューティング・ニューラルネットワーク(SCNN)の非決定論的学習手法を考案した。精度が僅かに低下するものの、長いビットストリームにより引き起こされるメモリ使用量増加を抑制できることがわかった。

研究成果の学術的意義や社会的意義
人工知能の実装に際し、これまで、厳密計算が不要であることに便乗した様々なデバイスを用いるアナログ計算手法が提案されてきた。しかし、主要なアプリケーションが、大規模言語モデルのように複雑化かつ大規模化するにつれ、アナログベースの計算基盤では太刀打ちできないことも明らかになってきた。今後、爆発的に普及することが予想される人工知能の実装手段としては、エネルギー効率を劇的に改善できるデジタル方式に期待がかかっている。本研究は、デジタル方式を維持しつつ、確率的計算を導入することにより、消費エネルギーを劇的に削減する手法として、有望である。

研究成果の概要（英文）：First, based on a new bisectional neural network (BNN), we devised a mechanism in which logical PE arrays are flexibly partitioned and mapped to physical computing units. By adjusting the shape and position of the logical PE array, we demonstrated the ability to spatially reconfigure functions. Temporal reconstruction ability in terms of precision could be achieved by adjusting the length of the SC bitstream. It is shown that the energy efficiency is better than other state-of-the-art approximate computing units. Second, we devised a non-deterministic learning method for stochastic computing neural networks (SCNN). It was found that the increase in memory usage caused by long bitstreams can be suppressed, although the accuracy is slightly degraded.

研究分野：計算機システム

キーワード：確率的デジタル計算　ニューラルネットワーク　非決定論的学習

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

様々な物理現象を演算・記憶能力に再解釈し、半導体の微細化限界を突破できるとする３方向の探求が盛んである。①**能動現象** リーク電流が少なく他素子を多数駆動できる金属酸化物等次世代半導体材料による低電力化手法。酸化済のため劣化に強く、複雑な演算・記憶能力を実現できる。②**受動現象** 微小記憶状態、スピン等量子状態、リザバー等非線形現象に対する干渉・観測に基づく計算。主要課題は、微細化技術(光メモリは原理的に困難)、安定性・可逆性・堅牢性、状態の干渉・観測に必要なエネルギーと演算時間の削減である。③**相互作用可能現象** 励起状態から安定状態に戻る過程をイジングモデルに写像し、組合せ最適化問題に対応。大規模化(相互作用範囲が限定される材料では困難)、状態の干渉・観測に必要なエネルギーと時間の削減、再現性の確保が難しい。大部分が、簡単な画像認識や小規模最適化問題に留まる中で、低電力性とスケーラビリティの両立が可能な有望技術が、スピントルク素子、メモリスタ、メムキャパシタ等記憶素子を格子状に配置し近似積和演算器を構成するインメモリ型アナログ積和演算である。抵抗素子は消費電力が大きく容量素子が有望、さらに従来型 Spike 密度表現は低速なため、発火時刻による確率的表現 Time-to-first-Spike(TTFS)と幅情報の併用が有力である。しかし、研究代表者は、以上の試みは、安定性と拡張性に優れたデジタル回路を超小型化する試みに対して、太刀打ちできないと予想している。近似計算の利用により可能となった超小型化技術を極め、本予想を実証的に明らかにする挑戦的研究が重要と考え、本研究構想に至った。

２．研究の目的

最先端 AI を含む次世代アプリケーションには、小型近似演算器に加え、キャッシュメモリが効かない離散データを効率よく扱えるシステムアーキテクチャが必須である。主記憶には、通常、１つのワード線に接続された多数のビットラインに１組のデータを記憶させる。本研究の当初の目的は、安定性と拡張性に優れたデジタル回路を利用しつつ、１つのビットラインに１組のデータを記憶させる手法を軸に、アドレス情報を付加した離散行列間の近似演算を高効率に実行する最適なアーキテクチャを探索し、キャッシュメモリに依存しない未踏の主記憶転置型・超小型離散行列シリアル演算機構を実現することであった。しかし、研究が進むにつれ、主記憶構造自体は改変しないほうが、様々な計算基盤への組み込みが容易になり、応用範囲も広がると考えた。そこで、具体的に、確率的再構成可能デジタルアクセラレータ（別名 DiaNet4）を従来型メモリに接続し、同等機能の実現を目指すことを新たな目的として再設定した。

３．研究の方法

新計算原理に基づくハードウェアと、最適化した AI アルゴリズムの組み合わせにより、超小型 AI デジタル計算基盤の研究を推進した。

**【1】3 値重みによるディープスパイキングニューラルネットワークの学習**

超小型 AI ハードウェアを開発する際に重要となるのは量子化である。トレーニング時の重み精度を下げることで、推論時のメモリ量と演算器を効果的に削減できる。シングルビット入力を使用するスパイキングニューラルネットワーク(SNN)でも、大量の浮動小数点重みのままでは、演算効率が大幅に低下する。本研究の目標は、**図１Ａ**に示すように、３値重みを用いて、SNN の計算時間を短縮することにある。具体的には、３値重みの SNN を直接学習するために、学習可能活性化パラメータと閾値、および、タイミング依存バックプロパゲーション（STDB）学習則を使う、Parametric Leaky-Integrate-and-Fire（PLIF）ニューロンを考案した。また、完全精度の重みと３値重みの量子化誤差を最小限に抑えるため、効率的な閾値ベース近似手法を考案した。さらに、**図１Ｂ**に示すように、より深い３値 SNN（TSNN）の学習のための、残差キャリブレーシ

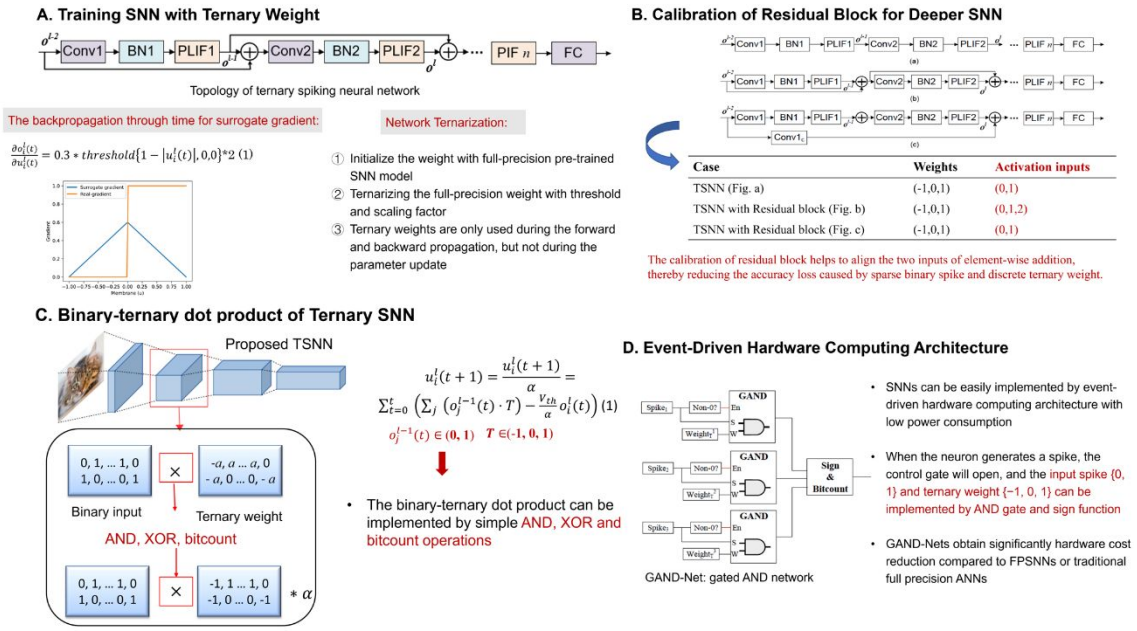


**図 1. 3値重みを使用したディープスパイキングニューラルネットワークの学習**

ョンを提案した。本フレームワークにより、SNN の内部状態は、**図１C**のように、単純な論理ゲートおよびビットカウンタを介する２値と３値のドット積に簡略化できた。既存の GXNOR-Net と同様、本提案の TSNN も計算ユニット制御ゲートを使用している。ただし、GXNOR-Net とは対照的に、TSNN のゲート信号は、スパイク/非スパイク状態を使用しており、演算ユニットは Gated AND Network （GAND-Net） と呼ぶ AND ゲートにより実装できる（**図１D**）。 本モデルを CIFAR-10、CIFAR-100、および、N-MNIST データセットを用いて評価した結果、より少ないタイムステップで同等の演算精度を達成できる目途が立った。

**【2】時空間再構成機能を備えた超小型確率計算ユニットの開発**

低精度の確率的計算 （SC） は、バイナリ計算よりも算術演算回路を小型化できるため、**図２A**に示すように、ニューラルネットワークの低電力実装に広く使用されている。ただし、エネルギー効率および柔軟性をより高めるために、詳細な検討が必要である。本研究では、二分ニューラルネットワーク （BNN） トポロジを有する再構成可能アーキテクチャに SC を導入し、時空間再構成機能を備えた超小型計算ユニットを考案した。本方式を採用した理由は、**図２B**に示すように、SC と BNN の相性が良いためである。BNN の単純な二分対称接続トポロジは、SC ベースニューロンの加算と乗算を効率的に表現できる。演算ユニットグループの形状および位置を調整することにより、空間（機能）的再構成が可能となる。また、各演算ユニットにおける SC ビットストリームの長さを調整することで、SC ベース演算の精度と実行時間をトレードオフし、時間 （精度） 的再構成も可能性となった。さらに、活性化機能は、**図２C**に示す有限状態マシン (FSM) を用いて効率的に実装できた。**図２D**に示すように、高精度回帰計算を実現するための最適 FSM 設計を探索した。

**【3】メモリ効率の高い確率的ニューラルネットワークのための非決定論的学習アプローチ**

ハードウェア視点での物量削減、および、アルゴリズム視点でのモデルパラメータ最適化により、モデルのさらなる小型化を行った。確率的計算ベースニューラルネットワーク （SCNN） は、通常、入力データとパラメータのエンコードに、極めて長いビットストリームを必要とし、**図３A**に示すように、ハードウェア量と計算時間が増加する。SCNN を改善する様々な提案があるものの、SCNN の学習に非決定性を導入して不要なビットを削減するような、パラメータ更新は考慮されていない。従来の SCNN における無駄を削減する、非決定論的学習アプローチを提案した。学習中のフォワードステップでは、ユニポーラまたはバイポーラのエンコード形式に対応して、

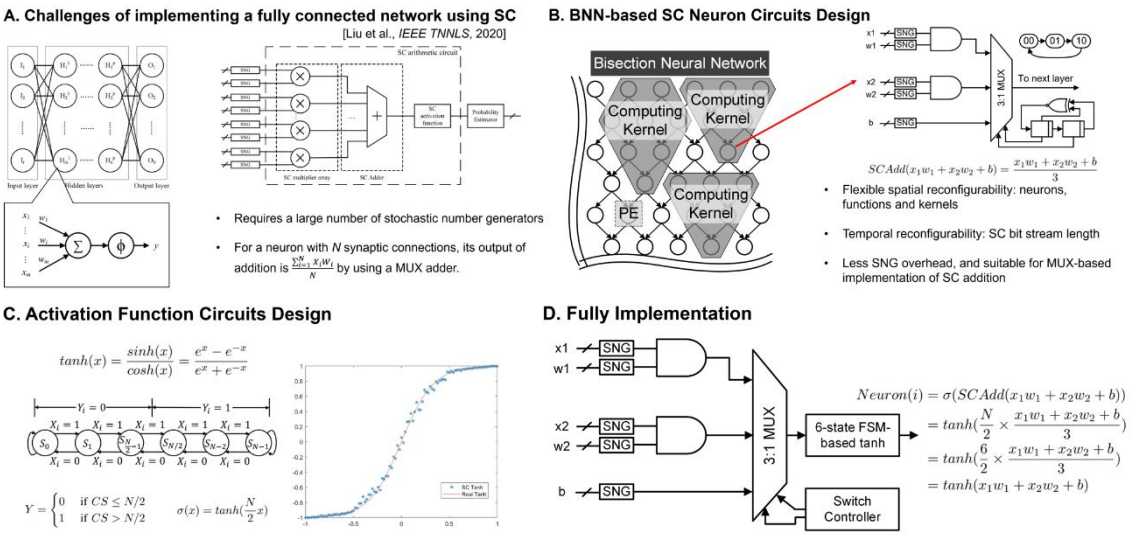


**図 2. 時間空間再構成機能を備えた超小型演算ユニット**

モデルパラメータを[0，1]または[−1，1]の範囲の確率に変換する。一様分布に従う乱数を導入し、**図３Ｂ**に示すように、上記確率に従い、パラメータを 1/4/8 ビットの確率数表現（Stochastic Number）に変換する。一方、バックワードステップでは、モデルパラメータを元の形式を用いて更新する。低精度量子化による発生する学習不安定性問題を解決するために、多重並列学習(MPTS）を提案した。MPTS は、**図３Ｃ**に示す投票機構を通じて、安定した学習を実現できる。**図３Ｄ**に示すように、完全結合の NN および MNIST データセットを用いて評価を行った。

## ４．研究成果

【2022 年度】

サロゲート勾配ベースの SNN および閾値ベースの３値重みパラダイムに着目した。GAND-Net と呼ぶ本提案は、バイナリスパイキングと離散３値重みの利点を組み合わせる構成である。探索の結果、SNN の内部状態は、イベント駆動制御信号と AND ゲートを介する２値と３値の内積に単純化された。この結果、GAND-Net は、精度と推論時間を維持しつつ、計算時間とメモリ量をともに削減できた。CIFAR-10、CIFAR-100、および NMNIST データセットを用いて評価した結果、より少ないタイムステップにより、87.42%、63.42%、および、98.43% の精度を達成した。GAND-Nets は、様々なデータセットに対して、３値 SNN アーキテクチャを適用した場合でも、一貫して有効性を発揮し、CIFAR-100 では 3.2%の精度向上と 25 倍の高速化が確認でき、バイナリ SNN モデルの性能を上回った。

【2023 年度】

時空間再構成可能超小型計算ユニットを実装した。 空間的再構成可能性は、二分ニューラルネットワークに基づき、時間的再構成可能性は、確率的コンピューティングに基づいている。 提案した演算ユニットは、様々関数計算に柔軟に対応できるのみならず、必要精度に合わせて確率数の長さを調整することも可能である。ハードウェア評価のために、Synopsys Design Compiler を使用し、45nm CMOS ライブラリを用いて、様々な深さの演算ユニットを合成した。評価の結果、16 ビット固定小数点バイナリ表現の BNN と比較して、同一動作周波数(833.33MHz)では、深さ 3、5、7 の SC BNN の消費電力を各々75.9%、71.6%、69.1%に削減できた。また、面積は、各々78.4%、77.6%、77.6%に削減できた。さらに、確率的ニューラルネットワーク（SCNN）のビットストリーム圧縮の検討も完了した。重みとバイアスは、学習プロセスのフォワード中では確率的エンコーディングに変換した。 バックワード中では、モデルパラメータは実数形式で更新した。評価の結果、提案する SCNN は、他の SCNN と比較して、より少ないビットストリームと投票機構でも安

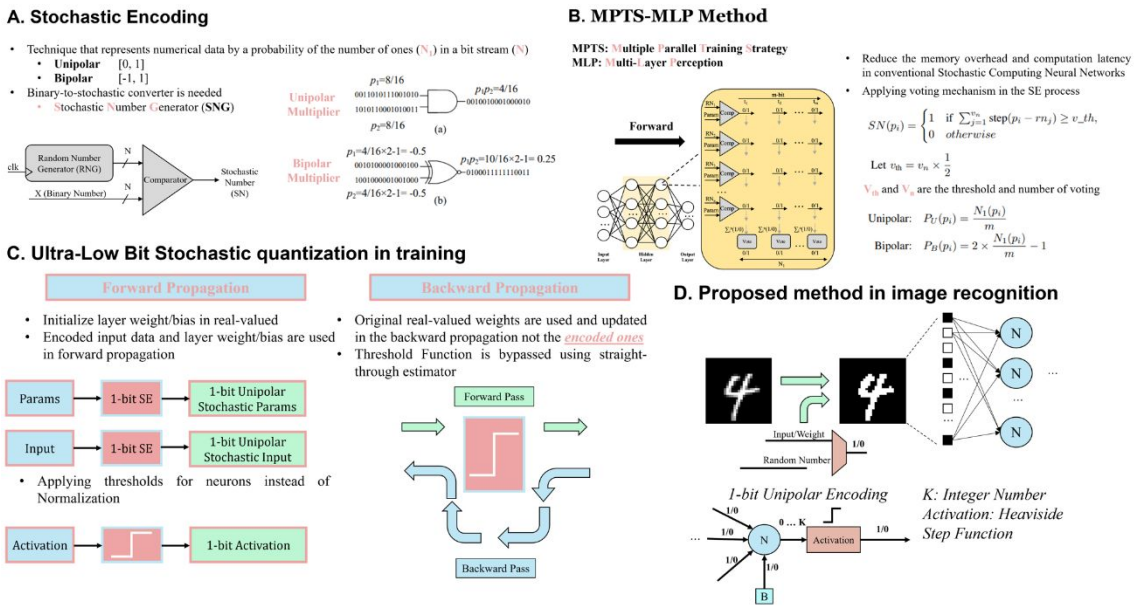


**図 3. メモリ効率の高い確率的ニューラルネットワークのための非決定論的学習アプローチ**

定した学習を実行でき、非決定性学習手法を利用することにより、ノイズのあるデータを扱う場合に、より信頼性の高い NN モデルを実現することを示した。精度は落ちるものの、元のモデルと比較して、メモリ量を削減できた。

【まとめ】

当初、主記憶構造の転置による、離散行列のシリアル演算効率化を目指していた。しかし、研究が進むにつれ、主記憶構造自体は改変せず、確率的再構成可能デジタルアクセラレータ、別名 DiaNet4 を従来型メモリに接続し、同等機能の実現を目指すこととした。

第１に、時空間再構成機能を備えた超小型計算ユニットを開発した。新しいバイセクション・ニューラルネットワーク(BNN)トポロジに基づき、論理 PE アレイが、ハードウェア上の物理計算ユニットに柔軟に分割写像される仕組みを考案した。論理 PE アレイの形状や位置を調整することにより、機能に関する「空間的再構成能力」を実証した。一方、BNN トポロジの利点を活用するために、物理計算ユニットに対して確率的コンピューティング(SC)ロジックを統合した。精度に関する「時間的再構成能力」は、SC ビットストリームの長さを調整することにより実現できた。以上の構成により、提案手法が、エネルギー効率の尺度において、他の最先端近似計算ユニットよりも優れていることを示した。

第２に、メモリ効率の高い確率的コンピューティング・ニューラルネットワーク(SCNN)の非決定論的学習手法を提案した。推論において、長いビットストリームを用いて厳密精度の NN を SCNN に変換する従来手法とは異なり、提案手法では、学習において、非決定論的計算の概念を導入し、長いビットストリームにより引き起こされるメモリ使用量の増加抑制を狙った。本目的を達成するために、学習時のフィードフォワードプロセスにおいて、NN パラメータを確率化し、確率に基づいて 1/4/8 ビットの確率数表現に変換する。これにより、SC のメモリ要件が大幅に削減できた。低ビット符号化により引き起こされる学習不安定性問題を軽減するために、学習中に、結果の安定性を向上させる複数並列学習戦略(MPTS)を提案した。評価の結果、アプリケーションの精度が僅かに低下するものの、SCNN のメモリ使用量を大幅に削減できることがわかった。以上のように、当初の計画以上に進展した。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3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<br>Man Wu, Yirong Kan, Renyuan Zhang, Yasuhiko Nakashima | ４．巻<br>ISBN:978-1-6654-5986-0 |
| ２．論文標題<br>GAND-Nets: Training Deep Spiking Neural Networks with Ternary Weights | ５．発行年<br>2022年 |
| ３．雑誌名<br>Proc. of 2022 IEEE 35th International System-on-Chip Conferen | ６．最初と最後の頁<br>1-6 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1109/SOCC56010.2022.9908132 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Zhu Guangxian、Kan Yirong、Zhang Renyuan、Nakashima Yasuhiko | ４．巻<br>1 |
| ２．論文標題<br>An Ultra-Compact Calculation Unit with Temporal-Spatial Re-configurability | ５．発行年<br>2023年 |
| ３．雑誌名<br>2023 21st IEEE Interregional NEWCAS Conference (NEWCAS) | ６．最初と最後の頁<br>1-5 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1109/NEWCAS57931.2023.10198176 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Golbabaei Babak、Zhu Guangxian、Kan Yirong、Zhang Renyuan、Nakashima Yasuhiko | ４．巻<br>1 |
| ２．論文標題<br>A Non-deterministic Training Approach for Memory-Efficient Stochastic Neural Networks | ５．発行年<br>2023年 |
| ３．雑誌名<br>IEEE International System-On-Chip Conference (SOCC2023) | ６．最初と最後の頁<br>05-08 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1109/SOCC58585.2023.10256838 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Sustainable Computing Architecture Lab.
http://archlab.naist.jp/
Sustainable Computing Architecture Lab.
https://www.youtube.com/channel/UC7rDo22e2SsucCY4FtpUWcg/videos

6．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究分担者 | ＺＨＡＮＧ　Ｒｅｎｙｕａｎ<br>(Zhang Renyuan)<br>(00709131) | 奈良先端科学技術大学院大学・先端科学技術研究科・客員教授<br>(14603) | |
| 研究分担者 | ＫＡＮ　ＹＩＲＯＮＧ<br>(Kan Yirong)<br>(50963732) | 奈良先端科学技術大学院大学・先端科学技術研究科・助教<br>(14603) | |

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 |
|---|---|