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　３ 年　 ５ 月 ３１ 日現在

機関番号：３７１１２
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１２３０

研究課題名（和文）どこでもAIに向けた省電力SRAMｾﾙｱﾚｲ多bit重みﾍﾞｸﾄﾙ機械学習識別器

研究課題名（英文）Low power multi-bit weight vector operated SRAM cell array machine learning classifier for an era of AI anywhere

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）　　3,500,000円

研究成果の概要（和文）：重みベクトルWrite用と内積和Read用のポートを分離し、7T-SRAMのRead用1-トランジスタの並列数・パルス幅の値で多bitの各電流値を調整し、ビット線短絡で加算した。この技術により「特徴ベクトル×多bit重みベクトルの内積和を、記憶データの破壊を回避しながら短絡ビット線電流に反映しワード線降圧を不要とし低電圧が可能」となった。量子化アルゴリズムは「特徴ベクトル行ごとに量子化することで、各ワード単位で正規化が可能になり、各bitが偏って1/0に丸められる確率を減らし、多bitの本来の精度を維持できる」ことが確認できた。結果、集団学習」を不要にし、大量のセルアレイを削減できた。

研究成果の学術的意義や社会的意義
本研究は「特徴ベクトル×1-bit重みベクトルの内積計算をSRAMセルアレイの読み出し動作で可能にする機械学習識別器に関するもので提案技術により「特徴ベクトル×多bit重みベクトルの内積和を短絡ビット線電流に反映できる。量子化アルゴリズムは「特徴ベクトル行ごとに量子化することで1/0に丸められる確率を減らし、多bitの本来の精度を維持できる」結果、コストを犠牲にしても必要だった「精度補償用の集団学習」を不要にし、大量のセルアレイを削減できる。これにより、本研究の目指す「どこでもAIに向けて必須の、省電力化機械学習識別器」に必要な、精度とコスト（消費電力、面積）のトレードオフの関係が改善される。

研究成果の概要（英文）：This research demonstrated that decoupling read-port from write-port is the key to a better stability for the machine learning operation because the product operations between the input vectors and weight value can be isolated form the storage nodes of the SRAM cells. Both (1) the number of multiple parallel connected transistors for the read-port and (2) the pulse width applied to the transistors can be used for the adjustment of amount of the product value without any stability issues. Because BL can be shorten without any stability issues. This removes the requirement for the suppressing WL voltage level, thus lower voltage operation is enabled. New algorithm for binary quantization reduced the instability of the weight learning curve thanks to regularizing the weight variation range.
The weights are regularized within the same WL unit. This reduced the accuracy degradation due to binary quantization error. This eliminated the requirements for the expensive ensemble learning.

研究分野：計算機システム

キーワード：省電力機械学習　SRAM内機械学習　メモリ内機械学習　量子化機械学習　１ビット機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

特徴ベクトル×重みベクトルの演算装置と外部データメモリの距離は遠く、頻繁なアクセスによる消費電力は大きい。逆に一連の処理がメモリアレイ内で完結できれば、桁違いの省電力化ができる。2015 年以降、「どこでも AI に向けて」メモリアレイ内機械学習の関連研究が活発化している。交差抵抗値などを重みに利用する研究もあるが抵抗バラツキなど課題も多く注目度は低かった。一方で、2016 年には標準の SRAM セルアレイ内で実現した手書き数字識別器が発表され[Zhang, N.Verma et al, IEEE Symposium on VLSI Circuit 2016]、大きな反響を呼んだ。しかし、以下の学術的課題が明らかになった。

(1)通常は、ワード線・ビット線を 1 本だけ選択した時の Read/Write 動作マージンを確保する構成になっており、全ワード線を一括 On するとセル内のデータが破壊される。ましてや異なる列のビット線を短絡できない)。低電圧化、微細化で増大する時空間バラツキへの脆弱性の課題も大きい。

(2)前記破壊を回避するために Zhang の論文ではワード線電位 Vi を電源電圧 VDD の 1/3 の 0.4V 以下で降圧している。

(3)1bit に丸められた重み誤差で精度が悪く、補償のために 18 列の Sign(+/-)結果にアレイ外部で重み(CN)を掛けて合算する集団学習が必要になる。1 個の識別に平均 18 列として 10 種の数字識別器には合計（10*(10-1)/2)*18=810 個の列が必要になり消費電力の点で犠牲が大きい。

☆☆ 上記問題を解決しなければ、「低電圧化と時空間バラツキ」の宿命的課題を背負ってのSRAM ｾﾙｱﾚｲ機械学習識別器の発展はなく、「どこでも AI に向けた AI×IoT センサーシステム」は実現不可能である。その解決技術を研究する学術的、社会的意味は大きいと言える。

２．研究の目的

本研究の独創的な点は「R/W ポートを分離し、データ破壊を回避しつつ、重み相当の電流能力を持つ R ビット線群を短絡させ、全ワード線を On にするだけで、多 bit 重みベクトルとの内積和を短絡ビット線電流に反映できる点」と「1-bit 量子化時に偏って 0/1 に丸められなくする独自のアルゴリズム」である。結果、精度補償用の大量の集団学習列アレイを不要にし、省電力化ができる点である。

本研究の目的は、SRAM セルアレイ機械学習識別器の電圧スケーリングの延命のために、機械学習特有アクセスが本質的に及ぼす課題を明確にし、上記した独創的な観点から課題を解決する手段を提案し実証することである。

３．研究の方法

【量子化誤差を削減するアルゴリズム】

学習モデルの精度は、重み w をセルの 1bit に量子化する時の丸め誤差で決まる。本研究の目的の 1 つは w 情報が偏って 0/1 に丸められないようにする事である。申請者が考案した「特徴ベクトルの行ごとに量子化する手法」が最も ”0 ”に偏って丸められるのを回避し ”1 ”が残っている。今後、この案に限定せず、何の量子化アルゴリズム、が同じ電力条件下で量子化誤差を最も低減できるのかを明らかにしていく。

【電力・面積効率を意識した多 bit 重み実現アーキテクチャー】

重みの多 bit 化を可能にした重要な発想は「R/W ポートの分離」である。個々のセルへの Write は、個々のビット線間分離が必然な一方で、Read 専用ビット線は列間で短絡できるので、例えば 3bit (22+21+20)化は、3 本のビット線の短絡だけで実現できる。各ﾋﾞｯﾄ線電流の重みは、1T セルの並列数で可変にできる。3bit 分(7T+8T+10T)は 6T セルの 4 個分の面積で実現できる。列ごとに A/D 変換して列間でデジタル加算する従来方式と比較して省面積・省電力にできる。各桁の重みは 6T の電源電位 VC や 1T のソース線電位 VR で独立に重みを設計可能である。

【アナログレベルの電流合成の回路設計とモンテカルロシミュレーション、識別マージンの影響解析】

0.4V に降圧したワード線のレベル以下で特徴ベクター v の量子化を行った場合の「時空間 Vt バラツキが識別マージンに及ぼす影響」を明らかにするために、時空間上ﾗﾝﾀﾞﾑに変動するビット線合成電流をモンテカルロシミュレーションで解析する。識別モデル間の電流分布のすそ野の距離「統計的なマージンに相当」を指標として、最長化可能なアルゴリズム・アーキテクチャーを見極める。

４．研究成果

初年度では、上記課題を以下の方法で解決する技術を開発した。（１）メモリセルを６Ｔセルではなく入出力分離が可能な８Ｔセルに変更することで、出力であるビット線をお互いに短絡でき、２ビットの場合には６Ｔセルよりもむしろ小さくなることを発見した。
（２）又、各セルのビット線はセルの記憶状態の安定性に影響を与えないので短絡可能である、時間軸に多ビット化する事が可能でワード線の電圧レベルの変調を回避できる。（３）重みの正負符号を８Ｔセルの８５％以上の面積を占有する６Ｔセルに相当する部分に記憶する。ベクトルの長さに相当する部分を残り１５％しか占有しない２Ｔセルで表現する。１５％のうち半分がトランジスタのチャンネル幅とすると、実質７％の部分がＮビット倍されるだけで実現できる。例え８ﾋﾞｯﾄにしても全体は４９％の増加に抑制できる。６Ｔセルに比較して、２０％以下の面積で実現できる。（４）単純にサイン符号器で（+1,-1)で２値かせずに、他が提案しているスケーリング、バッチノーマリゼーションの技術以外に、特徴ベクトル方向毎のスケーリングを提案し、従来のスケーリングで課題だった小さな値が丸められる問題を解決できた。

２年目では、初年度の結果と本研究の目的を踏まえ、以下の方法で解決する技術を開発した。（１）メモリセルを入出力分離が可能な８Ｔセルに変更することで、出力であるビッ ト線をお互いに短絡でき、２～５ビットの場合の６Ｔセル比での面積削減量を定量化した。（２）入力やアクティベーション情報の表現値として、電圧方向にア ナログ的に多ビットするのでなく時間軸方向に多ビット化する事の有効性を検証した。（３）ビットシフトと時分割処理の組み合わせで多ビット情報を扱うこと の有効性の検証を開始し、同一コラムやロウ方向毎の処理制限を開放する可能性を見極めている（４）アンサンブル学習による面積、電力オーバーヘッドの問題 を解決するためのパラメータ値の表現ビット量子化技術の研究として、空間的領域制限、絶対値分布値の領域制限を、同一コラムやロウ方向毎ではなく、ランダムな空間での制限方式の有効性を検証開始した。

最終年度のまとめとしては、「特徴ベクトル×1-bit 重みベクトルの内積和を、ビット線電流に反映できる」SRAM セルアレイの読み出し動作をそのまま用いた機械学習識別器に関する研究を実施した。主に取り組んだ研究課題として、「1-bit 重みが 1-bit(0/1)に丸められることによる不安定な学習と得られる精度の課題」と、「全ワード線を On する時のメモリセルに記憶された重みのデータ破壊の心配があり、ワード線を極度に降圧する必要があるため、低電圧化の課題がある」点に注目し、本研究では、重みベクトル Write 用と内積和 Read 用のポートを分離し、7T-SRAM の Read 用 1-トランジスタの並列数・ゲート・ソース間駆動電圧・パルス幅の値で多 bit の各電流値を調整し、ビット線短絡で加算することを提案し確認した結果、有効性を明らかにできた。この提案技術により「特徴ベクトル×多 bit 重みベクトルの内積和を、記憶データの破壊の危険性を回避しながら短絡ビット線電流に反映できる。また、データ破壊防止のためのワード線降圧が必要なくなるため低電圧が可能」となった。もう一方の課題に対して提案した量子化アルゴリズムは「特徴ベクトル行ごとに量子化することで、各ワード単位で正規化が可能になり、各 bit が偏って 1/0 に丸められる確率を減らし、多 bit の本来の精度を維持できる」ことが確認できた。結果、消費電力、面積のコストを犠牲にしても必要だった「精度補償用の集団学習」を不要にし、大量のセルアレイを削減できることが明らかになった。このことにより、本研究の目指す「どこでも AI に向けて必須の、省電力化機械学習識別器」に必要な、精度とコスト（消費電力、面積）のトレードオフの関係が改善されることが明らかになった。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　6件／うちオープンアクセス　4件）

１．著者名 Cheng-Xin Xue, Wei-Cheng Zhao,Tzu-Hsien Yang, Yi-Ju Chen, Hiroyuki Yamauchi,Meng-Fan Chang	４．巻 54
２．論文標題 A 28nm 320Kb TCAM Macro using Split-Controlled Single-Load 14T Cell and Triple Margin Voltage Sense Amplifier	５．発行年 2019年
３．雑誌名 IEEE Journal of Solid-State Circuits	６．最初と最後の頁 2743-2753
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/JSSC.2019.2915577	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 Xin Si;Win-San Khwa;Jia-Jing Chen;Jia-Fang Li;Xiaoyu Sun;Rui Liu;Shimeng Yu;Hiroyuki Yamauchi; Qiang Li ;Meng-Fan Chang	４．巻 66
２．論文標題 A Dual-Split 6T SRAM based Computing-in-Memory Unit-Macro with Fully Parallel Product-Sum Operation for Binarized DNN Edge Processors	５．発行年 2019年
３．雑誌名 IEEE Transactions on Circuits and Systems I: Regular Papers	６．最初と最後の頁 4171-4185
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCSI.2019.2928043	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 Jiazhen Xi and Hiroyuki Yamauchi	４．巻 8
２．論文標題 A Column Reduction Technique for an In-Memory Machine-Learning Classifier	５．発行年 2018年
３．雑誌名 International Journal of Machine Learning and Computing	６．最初と最後の頁 １２７－１３２
掲載論文のDOI（デジタルオブジェクト識別子） 10.18178/ijmlc.2018.8.2.675	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 －

１．著者名 Worawit Somha and Hiroyuki Yamauchi	４．巻 7
２．論文標題 A Segmentation Kernel Fitting Technique to Circumvent Extreme Deviation from Exponentially Descent Tail Distribution	５．発行年 2018年
３．雑誌名 International Journal of Electrical and Electronic Engineering & Telecommunications (IJEEET)	６．最初と最後の頁 114-118
掲載論文のDOI（デジタルオブジェクト識別子） 10.18178/ijeetc.7.3.114-118	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 Wei-Hao Chen, Hiroyuki Yamauchi, Meng-Fan Chang and wt.al,	4．巻 39
2．論文標題 A Dual-Split-Controlled 4P2N 6T SRAM in Monolithic 3D-ICs with Enhanced Read Speed and Cell Stability for IoT Applications	5．発行年 2018年
3．雑誌名 IEEE Electron Device Letters (EDL)	6．最初と最後の頁 1167-1170
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LED.2018.2850322	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 Hiroyuki Yamauchi, Worawit Somha	4．巻 1047
2．論文標題 An Error Reduction Technique in Richardson-Lucy Deconvolution Method	5．発行年 2018年
3．雑誌名 IOP Conf. Series: Journal of Physics: Conf. Series	6．最初と最後の頁 1-13
掲載論文のDOI（デジタルオブジェクト識別子） 10.1088/1742-6596/1047/1/012017	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 Jianzhen Xi and Hiroyuki Yamauchi	4．巻 Vo.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 (IJPRAI)	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 オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 鶴隆介 , セキ カテイ , 山内寛行
2．発表標題 SRAMアレイ内機械学習器のコラムイ削減手法
3．学会等名 2019 年度第72回電気・情報関係学会九州支部連合大会
4．発表年 2019年

１．発表者名
カテイ セキ，鶴　隆介，山内寛行

２．発表標題
Approximately Quantizing Algorithm for In-memory Machine Learning Classifier

３．学会等名
第17回情報科学技術フォーラム FIT2018

４．発表年
2018年

１．発表者名
Hiroyuki Yamauchi

２．発表標題
A Power Saving Techniques for Machine Learning Edge Computing: Toward an Era of AI Everywhere

３．学会等名
The 2nd International Conference on Electronics, Communications and Control Engineering (ICECC2019)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Hiroyuki Yamauchi

２．発表標題
Key Trials for a Technological Breakthrough to Realize an Ultra Energy-Efficient Machine Learning Computing

３．学会等名
The 3rd International Conference on Electronics, Communications and Control Engineering (ICECC2020)

４．発表年
2020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