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　 ６ 月 ２３ 日現在

機関番号：１４３０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１３８００

研究課題名（和文）メモリ内演算に基づく超低消費電力深層学習チップの開発

研究課題名（英文）Development of low power deep learning LSI based on in-memory computing

研究代表者
粟野　皓光（Awano, Hiromitsu）

京都大学・情報学研究科・准教授

研究者番号：１０７９９４４８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：メモリ内計算に適した機械学習アルゴリズムの開発と、ハードウェア実装を指向した性能評価に取り組んだ。具体的にはセルオートマトンと確率的データ構造の1種であるブルームフィルタを組み合わせた画像分類アルゴリズムを提案した。さらに65nmプロセスを想定した電力シミュレーションにより、既存手法と比較して推論精度を損なうことなく、50%の電力削減が可能であることを明らかにした。

研究成果の学術的意義や社会的意義
深層ニューラルネットワーク（DNN）が画像分類をはじめとして様々な領域で成果を挙げている。しかし、他を圧倒する性能と引き換えに、DNNの学習・推論に要するエネルギは膨大であり、DNNのエッジ応用を妨げる要因となっている。本研究では、リザバー計算に基づく反復法に依存しない機械学習アルゴリズムを提案し、電力削減効果を検証した。脱炭素が叫ばれる現代において、情報システムが消費する電力は膨大であり、その削減が強く求められている。本研究は知的コンピューティングの高効率化に向けて寄与する成果であると言える。

研究成果の概要（英文）：We developed a machine learning algorithm suitable for in-memory computation and evaluated its performance when implemented on silicon. Specifically, we proposed an image classification algorithm that combines cellular automata and Bloom filters, a probabilistic data structure. Furthermore, through power simulations assuming a 65-nm CMOS technology, we showed that the proposed algorithm can reduce power consumption by 50% without compromising inference accuracy compared to existing methods.

研究分野：集積回路

キーワード：リザバーコンピューティング　機械学習　低消費電力

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

深層ニューラルネットワーク（DNN）が画像分類を始めとした様々な領域で既存アルゴリズムを上回る成果を挙げている。しかし、他を圧倒する性能と引き換えに、DNN の学習・推論に要する計算コストは非常に大きく、DNN のエッジデバイス応用が広がらない要因となっている。そこで、誤差逆伝播などの反復法に依存しない、学習・推論が軽量な機械学習アルゴリズムの開発が求められている。

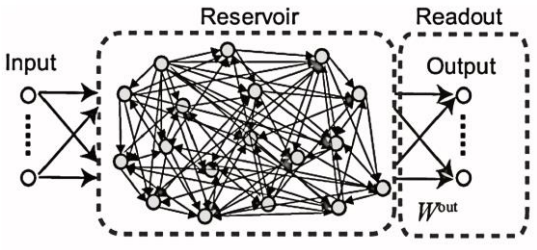
２．研究の目的

本研究では、確率的勾配法等の反復法に依存しない、機械学習手法を開発し、画像分類タスクにおいてその有効性を検証する。具体的にはリザバーコンピューティングに着目し、リザバーと後段のクラス分類器を、ディジタルハードウェア実装に適する様に改良することで、更なるエネルギ効率の向上を狙う。

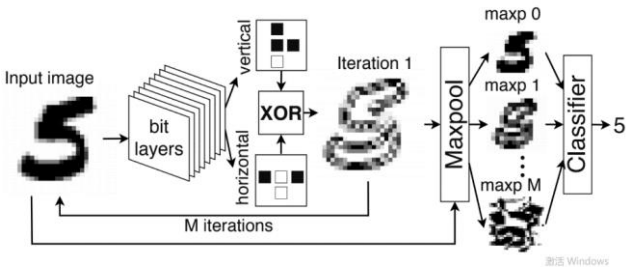
３．研究の方法

DNN は誤差逆伝播法によって、所望の入出力関係が実現されるようにネットワーク重みが調整される。しかし、誤差を逆伝播するためには、順伝播時の活性化値を全て保持しておく必要があり、層が深くなるにつれて学習に必要なメモリが爆発的に増大してしまうという問題がある。そこで、より簡易な学習が可能なネットワークとしてリザバーコンピューティング（RC）が注目を集めている。

RC の一種である Echo State Network（ESN）の構造を右図に示す。ESN では、リザバー内部のシナプス結合重みはランダムに初期化された後に固定され、学習では変化しない。リザバー内部と入出力層の結合重みは学習によって調整されるが、これは線形回帰問題と同等であるため、誤差逆伝播は必要とせず、非常に高速かつ省メモリな学習が可能となる。



ESN は再帰結合層を持つことから時系列データの取り扱いを念頭に置いている。一方、RC の考え方である、ランダムネットワークによって入力を高次元に写像し、学習が容易な軽量出力層で所望の値を取り出すという手法は画像認識等への応用も広がりつつある。右図に RC を利用した画像認識ネットワークを示す。このネットワークも特徴量候補を作り出すリザバーモジュールと、リザバーモジュールが作った特徴量から必要なものを抜き出して画像分類する識別モジュールから構成される。各モジュールの動作は以下の通りである。



[リザバーモジュール] : 入力画像における各ピクセルの輝度値は 8 ビットのバイナリ値に分解され、これを 8 チャンネルのバイナリ画像と見做してリザバーによって高次元空間に写像する。具体的にはバイナリ値を持つピクセルをセルオートマトン（CA）のルールに従って更新し、新しい画像を得る。CA のルールを k 回適用することで、同一の入力画像に対応する k 枚の相異なる画像を得ることが出来る。これらは MaxPooling によって解像度を落とした後に 1 次元ベクトルに変形され、各チャンネルのデータを 1 つのベクトルに結合して画像分類モジュールに送られる。

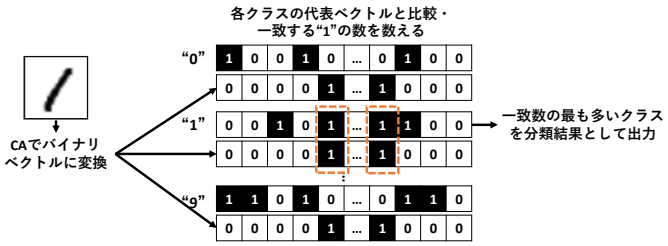
[画像分類モジュール] : 線形変換によってリザバーモジュールで作られたバイナリベクトルをクラス分類確率（logits）に変換する。

この手法（以下、ReCA と呼ぶ）は畳み込みニューラルネット（CNN）の計算で大部分を占める畳み込み演算を、CA に基づくリザバーで代替することで学習に要する計算コストの大幅な削減に成功している。一方、リザバーの出力する高次元バイナリベクトルをクラス分類確率に変換するために、画像分類モジュールには大規模な全結合層が必要となってしまい、結果としてモデルサイズは削減されないどころか増大してしまうという欠点を抱えている。

そこで、本研究では、Bloom Filter（BF）によって画像分類モジュールの全結合層を置き換えることで、モデルサイズを大幅に削減するとともに、反復法を必要としない高速な学習を実現す

る。

右図に提案手法のブロック図を示す。まず、入力画像をセルオートマトン（CA）によって高次元ベクトル表現に変換する。得られた高次元ベクトル表現は確率的データ構造の一種であるブルームフィルタ（BF）に送られ、各クラス（今回は MNIST を想定しているため“0”から“9”の 10 クラス）に対応する代表ベクトルと類似度を計算し、最も一致したクラスを分類結果として出力する。BF とは、入力された要素が集合のメンバである可能性があるか、あるいはメンバに含まれないかと効率的に調べることの出来るデータ構造である。BF の構成手順は以下の通りである。



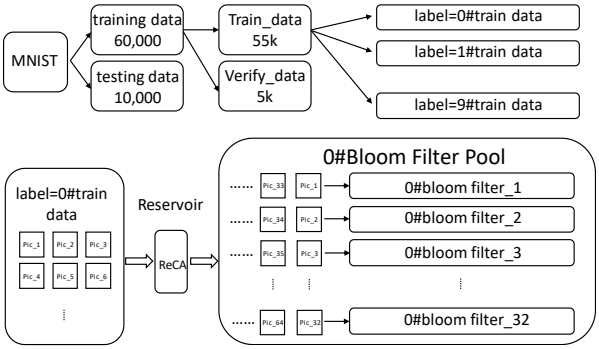
1. 全てのビットが“0”で初期化された m ビット配列を用意する
2. 特定の値（今回の例では 28×28 の値を持つ手書き文字）をハッシュ関数（今回の例では CA で構成されるリザバー）で m ビットの値に変換する
3. 2 で変換された値と、学習対象の BF との論理和を取り、結果を同じ BF に上書きする

ここで、手順 3 では論理和を計算していることから、既に BF に書かれている“1”の値は変化せずに、新しい入力に含まれる“1”が積み重ねられていくことに注意されたい。新しい学習された BF に対して、新しい入力が含まれるかを確認する手順は以下の様になる。

1. 学習時と同じハッシュ関数によって入力を m ビットの値に変換する
2. 変換された値で“1”が立っている位置で、同じく BF の値も“1”が立っていれば入力データは当該 BF に含まれる可能性がある

BF ではハッシュ値の衝突によって登録されていないデータも「含まれる」と判断してしまう可能性（偽陽性と呼ばれる）がある代わりに、非常に高い空間利用効率を実現できる。偽陽性は m が小さいほど、そして登録されるデータ数が多いほどに高くなる。そこで本研究では(1)畳み込み計算のデータフローを流用することで複数の BF の多数決によってクラス分類を行う Ensemble Bloom Filter（EBF）を提案し、学習データを複数の BF に分散して格納することで偽陽性を低下させる。

提案する EBF の学習法を右図に示す。“0”から“9”の各クラスには、それぞれ 32 個の BF が割り当てられており、これらの多数決でクラス分類を行う。BF の学習には 60k 枚ある学習画像のうち、55k 枚を使う。その手順は以下の通りである。



1. CA で画像を高次元に写像する
2. 写像された画像から 5×5 のパッチをストライド 3 で切り出す
3. 切り出された 256 個のパッチを 1 次元のバイナリベクトルに reshape し、BF に格納するデータを作る（1 データは 5kB）
4. 学習データを格納させる BF はラウンドロビン方式で決定する。具体的には 1 番目からはじめ、順次 1 つ隣の BF へと移動しながらデータを格納する。32 番目まで格納が終われば、再び 1 番目に戻って同様の手順を繰り返す

各クラスには 32 個の BF が割りついているが、BF の中にはクラス分類に寄与しないものも存在する。また、全ての BF を使って推論するためには、1.56MB ものメモリが必要となってしまうために効率が悪い。そこで、BF の学習で残しておいた 5k 枚の画像を利用して、不要な BF を枝刈りする。最終的には各クラスにつき 2〜4 個程度の BF を残し、残りを削除する。

４．研究成果

右表に実験結果を示す。提案手法は、同様に BF を用いている Bloom WiSARD と比較して 1/4 程度のメモリで同等以上の画像分類精度を達成出来ていることが分かる。

Method	Process	Memory (KB)	Accuracy(%)
Bloom WiSARD [Santiago et al., 2019]	+Distribution +Hash	819	91.50
Our work	+Distribution	1600	93.35
	+Distribution +Majority voting	800	89.12
		400	88.87
		200	88.40
	+Distribution +Majority voting +CA	**800**	**92.83**
		400	**92.20**
		200	**92.07**

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1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 Hiromitsu AWANO, Tadayuki ICHIHASHI, Makoto IKEDA	４．巻 E102.A
２．論文標題 An ASIC Crypto Processor for 254-Bit Prime-Field Pairing Featuring Programmable Arithmetic Core Optimized for Quadratic Extension Field	５．発行年 2019年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 56-64
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E102.A.56	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0件

〔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