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　４ 年　 ５ 月 ３０ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０２１８８

研究課題名（和文）撮像素子とアナログCNN回路の集積化により画像認識のエネルギーを1/1000倍に

研究課題名（英文）Integration of Image Sensor and Analog CNN Circuits Reducing Image Recognition Energy by Factor of 1/1000

研究代表者
高宮　真（Takamiya, Makoto）

東京大学・生産技術研究所・教授

研究者番号：２０４１９２６１

交付決定額（研究期間全体）：（直接経費）　　12,600,000円

研究成果の概要（和文）：深層畳み込みニューラルネットワーク(CNN)を用いた画像認識が高精度の画像認識を行う技術として、車の自動運転用カメラや監視カメラなどにおいて注目されている。CNNモデルの複雑化と画像認識に入力する画像の高解像度化に伴い、畳み込み演算量が年々、爆発的に増加しているため、低消費電力かつ低遅延時間のCNN処理を行うハードウエアが求められている。
そこで、本研究では低消費電力かつ低遅延時間のCNN処理を行う目的で、撮像素子とCNN演算回路を同一ICに集積化した「デジタルIn-Imager二次元畳み込みニューラルネットワークアクセラレータIC」を提案し動作原理を実証した。

研究成果の学術的意義や社会的意義
本研究成果「デジタルIn-Imager二次元畳み込みニューラルネットワークアクセラレータIC」により、将来、深層畳み込みニューラルネットワークを用いた高精度の画像認識を低消費電力かつ低遅延時間に実現可能になることが期待される。

研究成果の概要（英文）：With the widespread use of AI technology, ultra-low latency convolutional neural network (CNN) processing is highly demanded in fields that require real-time image classification such as autonomous driving and VR applications. This paper proposes an ultra-low-latency all-digital in-imager 2D binary convolutional neural network (II2D-BNN) accelerator for image classification. In II2D-BNN, multiply-accumulate operations (MACs) are processed inside the imager array parallelly in 2D, without extra latency for the row-by-row processing and data access with random access memories (RAMs). Convolution and sub-sampling operations using a 3 × 3 kernel are completed in only nine steps of batch-processing-in-2D regardless of image size using the II2D-BNN architecture, leading to over 88.5% reduction in computing latency compared with state-of-the-art architectures using batch-processing-in-1D.

研究分野：集積パワーマネジメント

キーワード：画像認識　畳み込みニューラルネットワーク　撮像素子　エネルギー

１．研究開始当初の背景

近年、畳み込みニューラルネットワーク（CNN）が高精度の画像認識を行う技術として注目されている。しかし、CNN モデルの複雑化と画像認識に入力する画像の高解像度化に伴い、畳み込み演算量が年々、爆発的に増加しているため、CNN 処理の遅延時間が増加し、リアルタイム性が要求される画像認識応用では問題となっている。例えば、車の自動運転に向けた画像認識の遅延時間は 170ms 以下が求められ、VR デバイスでは、50～100ms 以下の遅延時間が必要である。一方、画像認識で多用される CNN モデルの一つである VGG-Net では、169ms～4.3s の遅延時間が報告されており、リアルタイム画像認識の応用には適さない。図１（a）に従来 CNN アクセラレータの一次元畳み込み演算方法を示す。図１（b）に提案の二次元畳み込み演算方法を示す。図１（a）の従来では、演算ユニットとメモリとのデータアクセスが頻繁にあり、CNN の畳み込み演算を一次元的に一行ずつ処理しなければならないため、演算処理の遅延時間が画像解像度と共に増えていく。一方、図１（b）では、CNN の畳み込み演算を二次元で一括処理する In-imager 二次元畳み込み処理構造を提案し、画像認識の遅延時間が入力画像の解像度に関わらずに大幅に減少することができる。

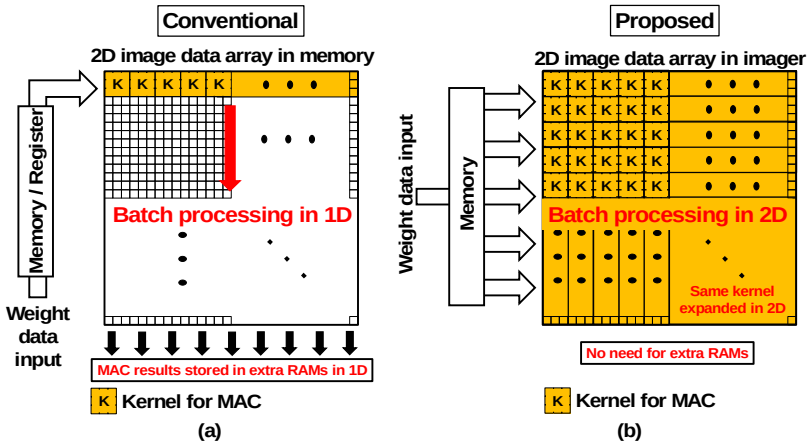

図１．（a）従来 CNN アクセラレータの一次元畳み込み演算方法と（b）本研究が提案した二次元畳み込み演算方法

２．研究の目的

本研究では、CNN 演算の遅延時間を大幅に減少することを目的として、世界初の In-imager 二次元畳み込み演算構造の CNN アクセラレータを提案する。本研究の目的では、提案した二次元演算構造をイメージャ回路の中に集積したアクセラレータを設計･試作･実測を行うことによって、提案した回路が撮像回路の内部でピクセル並列の二次元一括で畳み込み演算を行うことにより、畳み込み演算の遅延時間が入力画像サイズに依存せずに超低遅延で畳み込み演算を実現することを検証することである。

３．研究の方法

本研究では、リアルタイム画像認識に向けた高速な畳み込みニューラルネットワークを実現するために、デジタル演算回路をイメージャ回路の中に集積した In-imager 二次元畳み込みニューラルネットワークアクセラレータを提案した。MNIST の手書き数字画像認識を行う「デジタル In-Imager 二次元畳み込みニューラルネットワークアクセラレータ IC」を 180 nm 標準 CMOS プロセスを用いて設計し、試作を行った。本 IC ではピクセルと畳み込み演算回路が 30×30 の 2 次元アレーを構成している。提案回路を実証するために、実測・評判を行った。

４．研究成果

本研究では、従来の CNN アクセラレータに存在する一次元演算構造の壁を突破し、提案した In-imager 二次元畳み込みニューラルネットワークアクセラレータが入力画像の解像度に関わらず、高画像解像度でも超低演算遅延時間で CNN 演算ができることを検証した。主な成果は以下の通りである。

（１）世界初の In-imager 二次元畳み込み演算構造の提案

本研究では、提案の動作原理の実証を目的として、図２に示すような 30×30 の 2 次元アレーの「デジタル In-Imager 二次元畳み込みニューラルネットワークアクセラレータ IC」を設計し

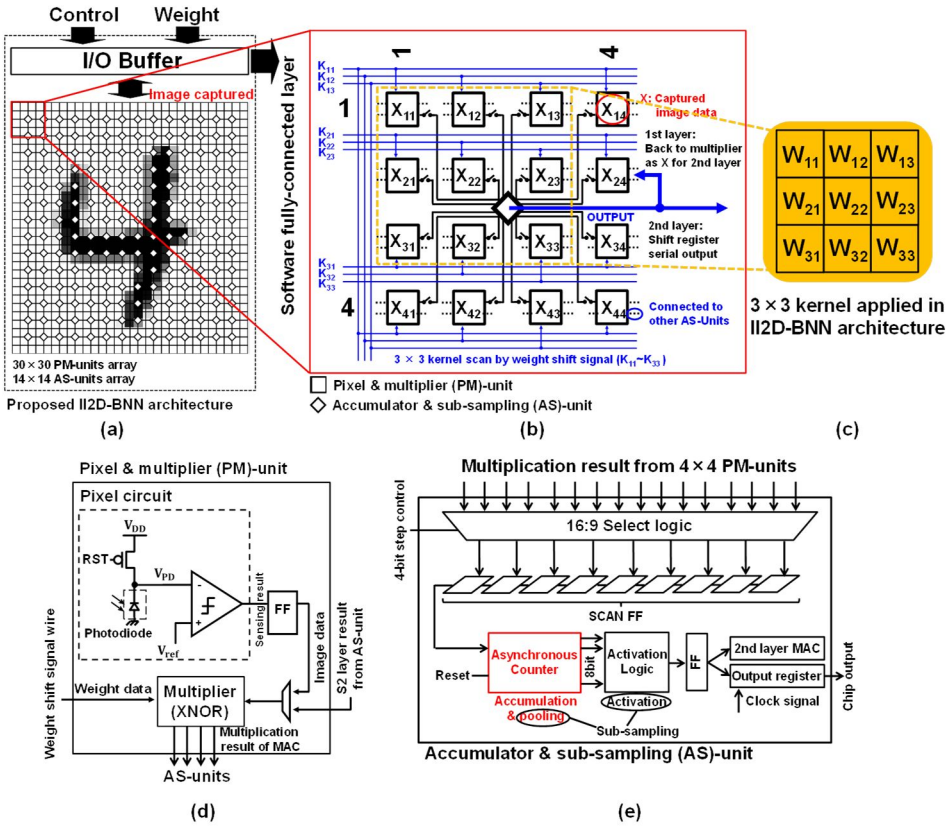

図２．(a)提案した回路の全体アーキテクチャ、(b) 提案回路内部の接続、(c)提案アクセラレータに応用する 3×3 カーネル、(d) 提案したピクセルと乗算器回路ユニット(PM-unit)の内部構造と (e) 提案した加算器とサブサンプリングユニット(AS-unit)の内部構造

た。提案したアクセラレータ回路が 30×30 のピクセルと乗算器回路ユニット(Pixel and multiplier unit (PM-unit))行列とその中に集積した 14×14 の加算器とサブサンプリングユニット(Accumulator and sub-sampling unit (AS-unit))回路からなり、画像データをキャプチャした直後に３×3 のカーネルを使って積和演算を行うことができる。本提案により、同一回路でピクセル並列の二次元一括で積和演算を行うことにより、畳み込み演算の遅延時間が入力画像サイズに依存せずに超低遅延で畳み込み演算を実現することができる。

（２）In-Imager 二次元畳み込みニューラルネットワークアクセラレータ IC の実証実験

180 nm 標準 CMOS プロセスを用いてした「デジタル In-Imager 二次元畳み込みニューラルネットワークアクセラレータ IC」のチップ写真を図３に示す。IC のチップサイズは 2.5mm 角である。図４にイメージャ回路の評価環境を示す。イメージャ回路の実測結果として、図５に 0 から 9 の手書き数字の入力画像とイメージャ回路がキャプチャした画像を示す、この実験によりイメージャ回路の動作チェックを確認した。一方、デジタル演算回路の動作検証を行った上、図 6 (a～e に提案アクセラレータの最高周波数、レイテンシ/カーネル、消費電力、そしてエネルギー効率（コアのみ）とエネルギー効率（I/O 含む）の電源電圧(V_{DD})依存をそれぞれ示す。この測定結果によって、提案アクセラレータでは電源電圧１V、クロック周波数 35.7MHz にて世界最短の 3.22μs/カーネルを達成し、従来研究より遅延時間を 80.5%削減した。また、提案アクセラレー

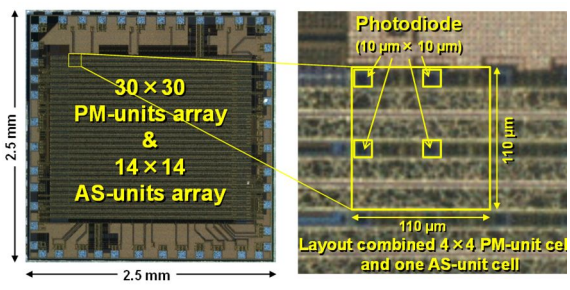

図３.試作したチップの写真

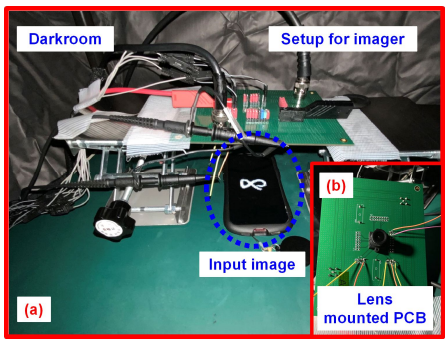

図４.イメージャ回路の評価環境

タが電源電圧 0.4V、クロック周波数 379kHz にて 5μW の低消費電力動作と、4.87 TOPS/W の高エネルギー効率動作を達成した。

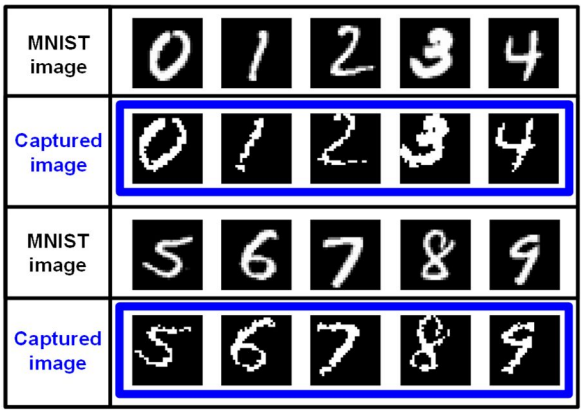

図５．0 から 9 の手書き数字の入力画像とイメージャ回路がキャプチャした画像

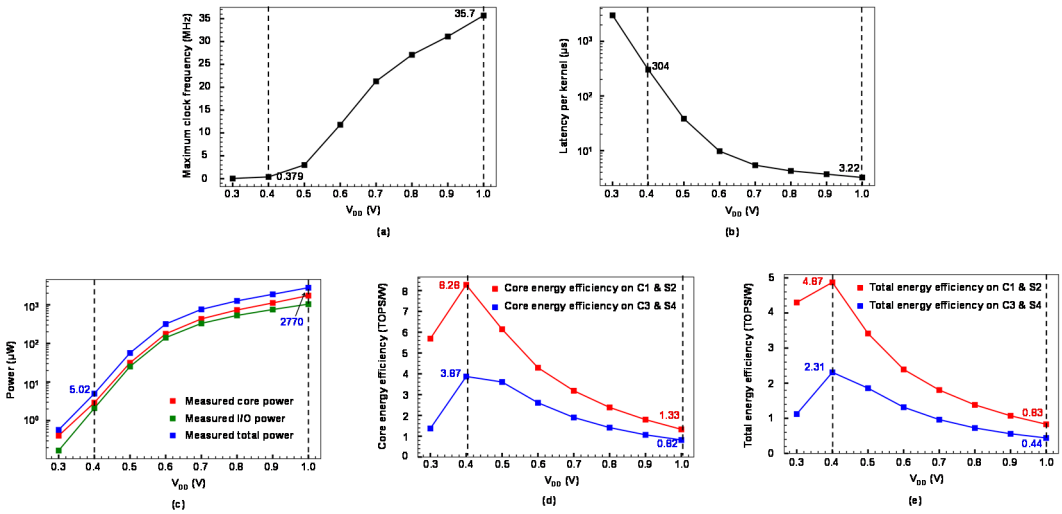

図６．デジタル In-Imager 二次元畳み込みニューラルネットワークアクセラレータ IC の(a)最高周波数、(b)レイテンシ／カーネル、(c)消費電力、(d)エネルギー効率（コアのみ）と(e)エネルギー効率(I/O 含む）の電源電圧(V_{DD})依存の実測結果

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 王 叡智，高宮 真
２．発表標題 超低遅延画像認識に向けたデジタルIn-Imager二次元畳み込みニューラルネットワークアクセラレータ
３．学会等名 電子情報通信学会，LSIとシステムのワークショップ，ポスターセッション 学生部門
４．発表年 2022年

１．発表者名 王 叡智，高宮 真
２．発表標題 高エネルギー効率のピクセル近傍2次元CNNアクセラレータの提案
３．学会等名 子情報通信学会ソサイエティ大会
４．発表年 2020年

１．発表者名 C. Wu and M. Takamiya
２．発表標題 Near-Pixel Binary Convolution Engine for Energy-Efficient Image Recognition
３．学会等名 電子情報通信学会総合大会
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