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　２ 年　 ７ 月　 ３ 日現在

機関番号：２６４０２
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００００８３

研究課題名（和文）深層学習エッジコンピューティング向け高効率組込みシステムの開発

研究課題名（英文）Highly Efficient Embedded Systems Architecture for Deep Learning Edge-Computing

研究代表者
密山　幸男（MITSUYAMA, YUKIO）

高知工科大学・システム工学群・准教授

研究者番号：８０３４６１８９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：深層学習エッジコンピューティングの高効率実行を可能にする再構成可能システムの実現を目指し、膨大な積和演算を効率良く実行することができる再構成可能アーキテクチャならびに、提案アーキテクチャに適した学習係数最適化手法を開発した。さらに、提案アーキテクチャのアプリケーション適用性を評価するため、市販のFPGAボードを用いた応用システムを構築した。ハードウェアアーキテクチャとソフトウェアアルゴリズムの両面から研究に取り組むことで、深層学習エッジコンピューティングを可能にする高効率組込みシステムの実現を目指した。

研究成果の学術的意義や社会的意義
これからのIoT社会において、IoTを構成するエッジデバイスの数とそこから得られるデータ量は爆発的に増加し、膨大なデータ転送と情報処理をサーバで行う計算モデルの限界が指摘されている。この問題を解決するため、本研究ではエッジデバイスからサーバへデータ送信する前に必要な処理を行う「エッジコンピューティング」の高効率化に取り組み、深層学習を利用したアプリケーションのリアルタイム実行を可能にする再構成可能アーキテクチャを提案した。提案アーキテクチャをFPGAの演算器ブロックやアクセラレータとして利用することで、高性能省電力なエッジコンピューティングを可能にする高効率プロセッサの実現に繋がると考える。

研究成果の概要（英文）：To realize highly efficient embedded systems for deep learning edge-computing, we have developed an efficient reconfigurable architecture for large-scale multiply-accumulation with enormous operands. We also proposed a weights quantization method with bit masking operations dedicated to the proposed reconfigurable architecture. Applied systems to demonstrate the effectiveness of our proposed reconfigurable architecture have also been developed.

研究分野：計算機アーキテクチャ

キーワード：再構成可能アーキテクチャ　深層学習　積和演算　アクセラレータ　組込みシステム

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

モノのインターネットと呼ばれる IoT の市場規模が拡大を続けており、IoT 社会が現実化しつつある。IoT 社会では、センサが取り付けられたさまざまなモノが互いにネットワークで接続され、社会生活を豊かにするこれまでにないサービス・アプリケーションの誕生が期待されている。IoT を構成するエッジデバイスからありとあらゆるデータが集められ、サーバに集約して情報処理を行うことが可能である。この膨大な集約データはビッグデータとも呼ばれ、これを解析することでさまざまな傾向を明らかにでき、幅広い応用が期待されている。ビッグデータ処理において近年注目されているのが機械学習である。人工知能技術のひとつであり、人間の持つ学習能力を計算機において実現するものである。ビッグデータに対して学習のための膨大な計算を行うことで、学習能力を獲得することができる。機械学習の応用範囲は、画像・音声認識のみならず、予測モデルの構築やシステムの異常検知など多岐にわたる。

その一方で、エッジデバイスの数は増加の一途を辿り、エッジデバイスからのデータトラヒックならびにサーバで処理する情報量や演算量が爆発的に増加することが課題となっている。これらのデータ転送ならびにデータ解析における遅延時間の問題は IoT 技術応用における大きな制約となっている。また、機械学習の一つである深層学習は、その高い認識精度からここ数年で急激に注目を集めているが、学習時はもとより学習結果に基づく認識処理においても膨大な計算が必要であることが課題となっており、一般の組込みプロセッサではリアルタイム処理は不可能である。このような背景から、エッジデバイスからデータを送信する前にデータを処理することでサーバへのデータ集中による諸問題を解決する「エッジコンピューティング」が注目されている。

２．研究の目的

本研究では、深層学習エッジコンピューティングの高効率実行を可能にする再構成可能システムの実現を目的としている。膨大な積和演算を効率良く実行することができる再構成可能アーキテクチャの研究に加えて、提案アーキテクチャに適した学習係数最適化手法についても研究を行う。また、提案する再構成アーキテクチャと最適化手法の有効性を示すため、実アプリケーションによる実証実験が可能な応用システムも構築する。ハードウェアアーキテクチャとソフトウェアアルゴリズムの両面から研究に取り組むことで、深層学習エッジコンピューティングを可能にする高効率組込みシステムの実現を目指す。

３．研究の方法

本研究目的を達成するため、以下に挙げる研究課題に取り組む。

(1) 再構成可能アーキテクチャの開発：
　任意の大規模積和演算の高効率処理を目指し、複数種類のカウンタで構成されるカウンタツリーを基本構成要素とする再構成可能アーキテクチャを開発する。チップ試作には 65nm SOTB プロセスを用いる。

(2) 学習係数最適化手法の提案：
　提案する再構成可能アーキテクチャによる高効率演算を可能にするため、提案アーキテクチャに適した学習係数の量子化アルゴリズムを開発する。

(3) 応用システムの開発：
　開発したアーキテクチャとアルゴリズムの有効性を、機械学習アプリケーションを用いて評価することができる評価用応用システムを構築する。

４．研究成果

4.1 再構成可能アーキテクチャの開発

深層学習やサポートベクタマシン（SVM）といった機械学習では特に大規模な積和演算の高効率実行が求められている。そこで、エッジデバイス上での機械学習に基づく識別処理を想定し、機械学習/識別処理に頻出する大規模積和演算などの高効率実行を可能にする多オペランド加算器を開発した。

6 オペランド積和演算を例に提案回路の概要を述べる。6 オペランド積和演算器は、図 1 に示すように 6 個の乗算器と 5 個の加算器で構成することができる（単純積和演算器）。これに対して提案する積和演算の構成は、図 2 に示すように部分積加算部に多段カウンタを基本要素とするアレイ構造を採用し、多数のオペランド（部分積）を毎サイクル足し合わせることができる。さらにキャリー伝搬加算を最終段の 1 回のみとすることで、プロセッサでの処理に対してレイテンシを大幅に削減し、高い性能が期待できる。また、プロセッサや GPU では演算ビット幅が固定であることに対して、提案する多オペランド加算器は演算ビット幅とオペランド数を可変とするプログラマビリティを有しているため、アプリケーションに合わせて回路リソースを最大限に使用することができる。これにより、プログラマブルな構造と ASIC に匹敵する演算効率の両立を目指している。

次に、提案再構成可能アーキテクチャに基づく多オペランド加算器を部分積加算部に搭載する大規模積和演算回路のテストチップを作成し、実測に基づく消費エネルギーの評価を行った。テストチップは、216 個の 12 ビットオペランドを入力とする大規模積和演算を実行することができる。まず、図 1 に示す単純積和演算器と比較した結果、論理合成による見積もり値では、演

算 1 回当たりの消費エネルギーは約 44%となり、十分な優位性を示した。次に、図 3 に示す評価ボードを用いてテストチップの電流値を測定し、消費エネルギーを評価したところ、論理合成結果と比較して約 7.4 倍の大きさになった。この原因として、テストチップの回路構成では組合せ回路の規模が大きく、膨大なグリッチが発生していることが考えられる。これは、提案回路内で適切にパイプラインレジスタを挿入することで解消され、消費エネルギーを大きく削減できるものである。このことから、提案アーキテクチャは、プログラマブルな構造と ASIC に匹敵する演算効率の両立を実現できると考えられる。

最後に、市販 FPGA とも比較評価を行った。テストチップと同じく 65nm プロセスで製造された Altera 社 StratixIII FPGA を対象とし、図 4 に示すカスタムボードを用いて測定した。テストチップと同じ規模の積和演算回路を実装して動作中の電流値を測定した結果、市販 FPGA はリーク電流が消費電流に占める割合がテストチップに対して 1 桁以上大きく、比較にならないほど消費エネルギーが大きかった。そこで、リーク電流を除いて比較したところ、100MHz 動作時においてテストチップは市販 FPGA の約 19%の消費エネルギーで演算を実行できることがわかった。

以上のことから、提案アーキテクチャの有効性を実証することができた。

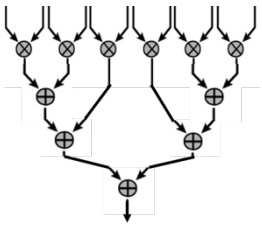

図1 単純積和演算器構造による
6オペランド積和演算の実行

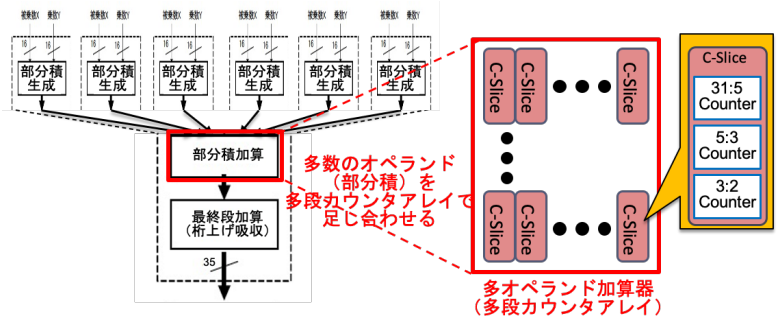

図2 提案加算器による6オペランド積和演算の実行

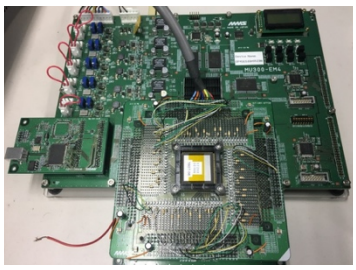

図3 テストチップ測定

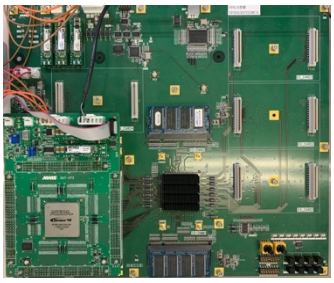

図4 StratixIII評価用カスタムボード

4.2 学習係数最適化手法の開発

4.1 章で述べた提案アーキテクチャは、アレイ構造を採用した再構成可能アーキテクチャであるため、使用リソースは係数（乗数）の各桁（ビット）が“1”を取り得る数に比例して大きくなる。すなわち、乗数の特定のビットを“0”でマスクすることができれば、同じビット幅の演算でも回路規模や消費電力を削減することができる。言い換えれば、見かけのビット幅を大きくすることができるため、機械学習の識別処理における精度向上が期待できると考えた。

そこで、深層学習のハードウェア実装で不可欠な学習係数の量子化処理において、単純にビット幅を削減するのではなく、大きなビット幅をとりながら特定のビットをマスクする手法について研究を進めてきた。提案量子化手法の概要を図 5 に示す。なお、量子化手順として、単精度浮動小数点演算で学習を完了した後、学習係数の量子化・ビットマスクを適用した。また、畳み込み層が複数ある場合は、それぞれの層で最適な量子化ビット数とマスクパターンを適用した。

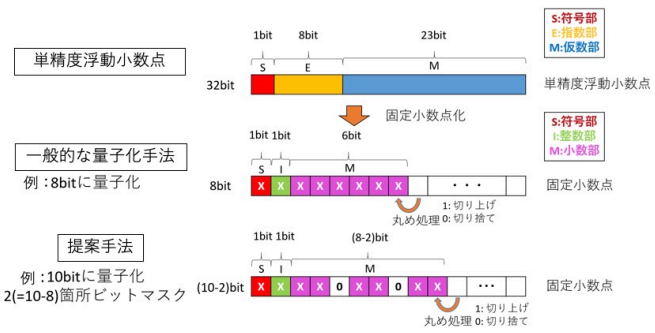

図5 提案手法の概要

表 1 に示すネットワーク構成とデータセットを用いて提案手法の有効性を評価した。評価結果を表 2 に示す。従来手法（単純量子化）（表 2 における「基準」）では認識精度が低かったビット数でも、提案手法によって認識精度が大幅に向上していることがわかる。このことから、提案アーキテクチャと学習係数最適化手法を組み合わせることで、単純量子化より少ない有効ビット数で高い識別精度が得られ、回路規模と消費電力の削減が可能であることがわかった。

表1 ネットワーク構成とデータセット

ネットワーク	LeNet_caffe	Cifar10_quick	Cifar100
データセット	Mnist	Cifar10	Cifar100
畳み込み層	2層	3層	3層
層構成	入力層 畳み込み層 プーリング層 畳み込み層 プーリング層 全結合層 出力層	入力層 畳み込み層 プーリング層 畳み込み層 プーリング層 畳み込み層 プーリング層 全結合層 出力層	入力層 畳み込み層 畳み込み層 畳み込み層 プーリング層 全結合層 出力層

表2 量子化パターンと認識率

ビット数	畳み込み層1	畳み込み層2	畳み込み層3	認識精度	
LeNet_caffe					
3bit	XX.X	XX.X		0.114	←基準
	X0.0XX	X0.00XX		0.962	←最高値
	X0.0XX	X0.00X0X		0.847	←最低値
4bit	XX.XX	XX.XX		0.232	
	X0.0XXX	X0.00XXX		0.976	
	X0.0XXX	X0.00X0XX		0.948	
Cifar10_quick					
4bit	XX.XX	XX.XX	XX.XX	0.101	
	X0.00XXX	X0.000XXX	X0.0000XXX	0.706	
	X0.00XXX	X0.000XX0X	X0.0000XX0X	0.560	
5bit	XX.XXX	XX.XXX	XX.XXX	0.101	
	X0.00XXXX	X0.000XXXX	X0.0000XXXX	0.737	
	X0.00XXXX	X0.000XXX0X	X0.000XXX0X	0.644	
Cifar100					
5bit	XX.XXX	XX.XXX	XX.XXX	0.028	
	XX.XXX	X0.0XXXX	X0.00XXXX	0.236	
	XX.XXX	X0.X0X0XX	X0.00XXXX	0.059	
6bit	XX.XXXX	XX.XXXX	XX.XXXX	0.216	
	XX.XXXX	X0.XXXXXX	X0.00XXXXX	0.351	
	XX.XXXX	X0.X0XXXX	X0.00XXXXX	0.130	

4.3 評価用応用システムの構築

提案する多オペランド加算器ブロックのアプリケーション適用性を評価するため、市販のFPGA ボードを用いて機械学習アルゴリズム（SVM）による人物検出処理のデモシステムの構築に取り組んだ（図 6）。多オペランド加算器は入出力数が膨大であるため、前述の試作チップを用いたアプリケーションデモが実施できない。そのため、提案する多オペランド加算器を市販 FPGA のプログラマブルロジック上に実装し、その他の処理はプロセッサコアなどに実装する。

実アプリケーション適用性をさらに評価するためのプラットフォームとして、機械学習アルゴリズムを用いた信号、歩行者、障害物、道路表示検出による自律走行システムを構築した。提案システムは、古典的な画像処理アルゴリズムに基づく走行制御系と、機械学習アルゴリズムによる物体検出系で構成される。特に物体検出系については、提案アーキテクチャに基づくテストチップを使った実装を想定しており、走行制御系と物体検出系を分けてそれぞれ個別の FPGA ボードで構成している。図 7 に自律走行車の全体システム構成を示す。また、図 8 に走行制御系ボードのプログラマブル SoC 上に実装した回路の構成を示す。物体検出系ボードにおいて機械学習アルゴリズムを用いて検出・認識した情報は、走行制御系ボードに GPIO を介して送信され、その後の動作制御は走行制御系が担う。研究実施期間中に物体検出系に試作チップを使用することはできなかったが、プログラマブル SoC 上に機械学習アルゴリズムに基づく信号検出処理を実装し、ミニチュア道路において車線を逸脱することなく安定して走行できることを確認した。

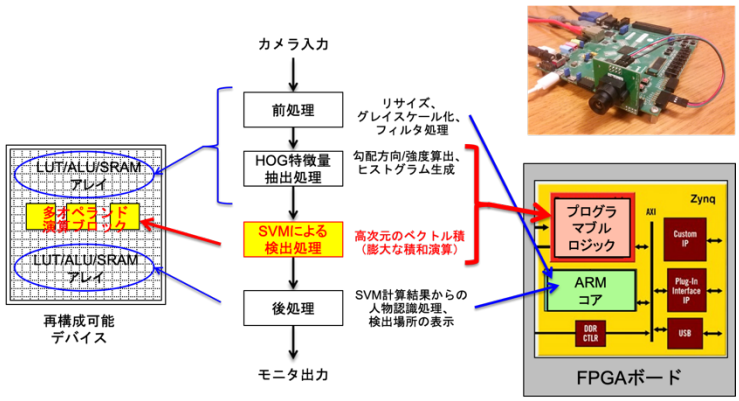

図6 機械学習（SVM）を用いた人物検出アプリケーションデモシステムの構成

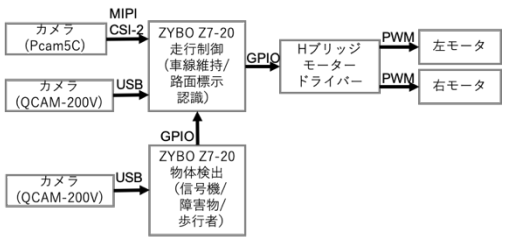

図7 全体システム構成

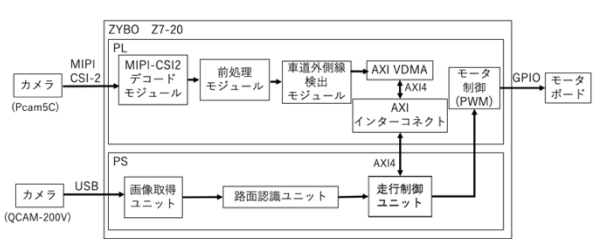

図8 走行制御系の回路構成

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 T. Tanaka, I. Ikeno, R. Tsuruoka, T. Kuchiba, W. Liao, and Y. Mitsuyama
２．発表標題 Development of Autonomous Driving System using Programmable SoCs
３．学会等名 International Conference on Field-Programmable Technology（国際学会）
４．発表年 2019年

１．発表者名 青野 遼, 長原 拓巳, 田中 知成, 池野 樹, 廖 望, 密山 幸男
２．発表標題 プログラマブルSoCを用いた小型自律走行車の構成検討
３．学会等名 電子情報通信学会 リコンフィギャラブルシステム研究会
４．発表年 2020年

１．発表者名 田中 知成, 池野 樹, 鶴岡 陸, 口羽 匠, 廖 望, 密山 幸男
２．発表標題 プログラマブルSoCを用いた自動運転システムの構成検討
３．学会等名 電気関係学会四国支部連合大会
４．発表年 2019年

１．発表者名 熊井 遼太, 密山 幸男
２．発表標題 プログラマブルSoCを用いたリアルタイム物体検出処理の実装
３．学会等名 電気関係学会四国支部連合大会
４．発表年 2019年

１．発表者名
兼本 一生，岡林 由真，風谷 亮太，密山 幸男

２．発表標題
多オペランド積和演算の高効率化に関する一検討

３．学会等名
電気関係学会四国支部連合大会

４．発表年
2019年

１．発表者名
熊井 遼太，和田 征也，密山 幸男

２．発表標題
高位合成系による人検出処理のFPGA実装と評価

３．学会等名
電気関係学会四国支部連合大会

４．発表年
2018年

１．発表者名
氏原 収悟，密山 幸男

２．発表標題
畳込みニューラルネットワーク向け重み量子化手法の検討

３．学会等名
情報処理学会 システムとLSIの設計技術研究会

４．発表年
2018年

１．発表者名
高野 雅之，熊井 遼太，毛利 真崇，小松 達也，密山 幸男

２．発表標題
高位合成系による人検出処理のFPGA実装

３．学会等名
電気関係学会四国支部連合大会

４．発表年
2017年

１．発表者名 氏原 収悟，密山 幸男
２．発表標題 深層学習を用いた画像認識処理における重み量子化のための評価環境構築
３．学会等名 電気関係学会四国支部連合大会
４．発表年 2017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