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　 ５ 月　 ３ 日現在

機関番号：２５３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2023
課題番号：１８Ｋ０４２８７

研究課題名（和文）画像認識処理向け動的分散並列パイプライン機構の研究

研究課題名（英文）A study of dynamic distributed parallel pipeline processing

研究代表者
有本　和民（Arimoto, Kazutami）

岡山県立大学・情報工学部・特任教授

研究者番号：１０５０１２２３

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：FPGAに適した組込み用ニューラルネットアーキテクチャを検討し、同FPGAに推論処理アクセラレータを実装する。FPGAへの実装方式の違いによる処理スループットと実行時消費電力が変化することを検証し、エッジ側でデータ収集・分析する低消費電力・低レイテンシのAIシステムシステム性能を評価した。実装にはXilinx ZCU102開発キット環境を使用し、センシングデータ用ニューラルネットを検討し、FPGAへの実装方式について性能と電力の比較調査を行った。回路実装時の並列パイプライン化によって処理性能を10倍以上に向上させることを確認し、枝刈りと並列パイプライン化との親和性が高い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
サイバーフィジカルシステムがクラウドベースから、エッジ・クラウド協調型へシフトする。エッジデバイスで低消費電力・低レイテンシのニューラルネットワーク推論処理を実行し、高電力効率のFPGAデバイスでの推論処理アクセラレータによるHWフォーム基盤技術が強く求められる。
IoT の普及により、多種多様なセンシングデータの収集や分析の重要性が高まり、センシングデータから有為な知見を得るためにAIが利用される。将来的に10億台以上にまで増加すると予想され、エッジ側でデータ収集・分析する低消費電力・低レイテンシのAIシステムとして、処理性能を10倍以上に向上させることを示した意義は大きい。

研究成果の概要（英文）：We studied an embedded neural network architecture suitable for FPGA, and implement an inference processing accelerator on the FPGA. We verified the changes in processing throughput and runtime power consumption due to differences in FPGA implementation methods, and evaluated the system performance of a low power consumption and low latency AI system that collects and analyzes data on the edge side.
For implementation, we used the Xilinx ZCU102 development kit environment, considered an embedded neural network for sensing data, and conducted a comparative investigation of performance and power regarding implementation methods on FPGA. We confirmed that parallel pipelining during circuit implementation improves processing performance by more than 10 times, and revealed that pruning and parallel pipelining are highly compatible.

研究分野：組込みシステム

キーワード：組込みシステム　FPGA　省電力回路　ニューラルネットワーク　エッジコンピューティング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

サイバーフィジカルワールドの発展と、IoT (Internet of Things) の普及にともない、多種多様なセンシングデータの収集や、収集したデータの機械学習等による精緻な分析の重要性がますます高まってきている。将来的に 10 億台以上にまで増加する（トリリオンセンサ時代とも呼称されている）と予想される IoT エッジデバイスの普及を考慮すれば、ネットワーク通信によってクラウド側でデータを収集・分析するよりも、エッジ側でデータ収集・分析する方がより低消費電力・低レイテンシの AI システム実現につながると考えられる[1]。ここでエッジ側での AI 実装を考える場合､従来の単一のセンサデバイスを用いたデータ収集･分析から、アレイ型などの複数個のセンサデバイスやを複数異種のセンサデバイスを用いた高精度のデータ収集・分析（センサフュージョン）への要求が高まってきており、本来は軽量であるはずの 1 次元時系列データ向け AI モデルが将来的には大規模化・大容量化していく（エッジヘビーコンピューティングへ移行していく）ことが予想されている。そのため、エッジ AI システムの低消費電力化・高電力効率化技術の確立は喫緊の課題となっている。

２．研究の目的

本研究は、エッジデバイスで低消費電力・低レイテンシのニューラルネットワーク (NN) 推論処理を実行するために、高い電力効率の FPGA デバイスを使った推論処理アクセラレータとそのハードウェアプラットフォーム基盤技術を確立することを目的とする。このうち本報告書では FPGA ハードウェアに適した組み込み用 NN アーキテクチャを検討し、同 FPGA に推論処理アクセラレータを実装する。FPGA への実装方式の違いによって処理スループットと実行時消費電力が変化することを検証し、アプリケーションを想定したシステム性能を評価することで、組み込み用 NN アーキテクチャの有用性を示す。加えて、エッジ AI システムの低消費電力化のために需要な、枝刈り技術との親和性についても検証する。

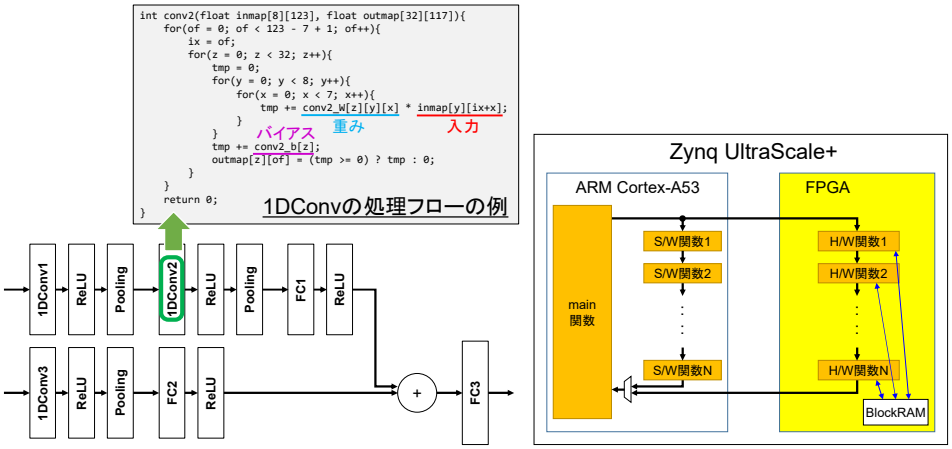
３．研究の方法

本研究では、推論処理アクセラレータとして Xilinx ZCU102 開発キット環境を使用し、1 次元時系列データ（センシングデータ）用の NN モデルについて組み込み用アーキテクチャを検討し、FPGA への実装方式について性能と電力の比較調査を行う。

４．研究成果

NN 推論モデルとその FPGA 実装のイメージを図 1 に示す。

図 1(a)では 1 次元畳み込み層 (1-Dimensional Convolution layer: 1DConv) と全結合層 (Full Connect layer: FC) を含む NN モデルの例を示しており、これはセンシングデータを入力とするモデルを想定したものである。ここで、1DConv 層に着目すると、1DConv 層は層への入力データと学習済みモデル (重みパラメータ) との畳み込み計算の繰り返しで構成されていることが分かる。ハードウェア実装を考える場合、層への入力データはバッファメモリから、また、学習済みモデルはメモリプールからそれぞれ供給され、層での計算結果が再びバッファメモリに書き戻される回路となる。バッファメモリやメモリプールは FPGA 内蔵 SRAM（BlockRAM : BRAM）を使って実装される。FC 層も 1DConv 層と同様に、計算部の繰り返し計算処理と、計算部と記憶部 (メモリ) とのデータ転送によって構成されており、FPGA ハードウェアに NN 推論モデルを実装する場合は、メモリの転送速度と計算回路の処理速度が全体の推論性能を決める重要な要因となる。



(a) NN構成の例　　(b) FPGA実装イメージ

図 1　ニューラルネットワークモデルとその FPGA 実装

図 1(b)ではひとつのマイクロコントローラ (MCU) コアに NN モデルをソフトウェアとして実装する場合と FPGA ハードウェアに NN モデルを実装する場合のイメージが描かれている。画像認識に関してはいくつかの研究実績があるが、本報告では、一例として、触覚センサ出力での画像認識による紙識別を想定した NN モデルを例にして述べる（図 2 参照）

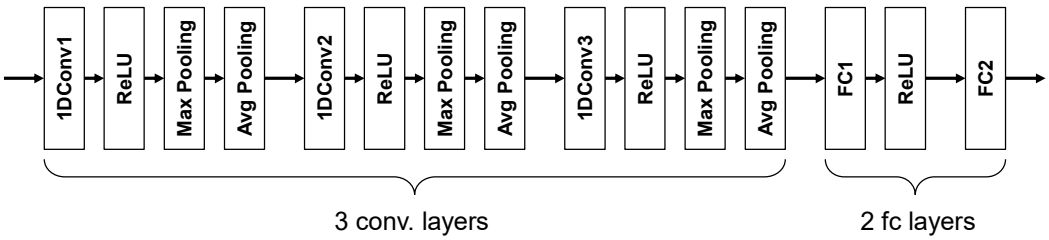
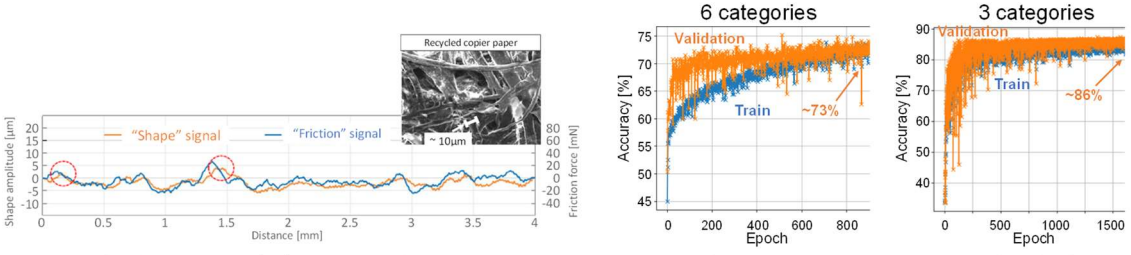


図 2　触覚センサによる紙識別用の NN モデル

1 次元畳み込み層が 3 層と全結合層が 2 層の、合計で 5 層構成の NN モデルであり、畳み込みの計算結果に対して最大値プーリングと平均値プーリングの両方を実行している。この NN モデルを用いて図 3(a)の触覚センシングデータを入力とした場合の学習結果を図 3(b)に示す。6 種類の紙を用いる場合の紙識別精度が約 73%、3 種類の紙を用いる場合の紙識別精度が約 86%となっている。この NN モデルを題材に、FPGA への推論アクセラレータの実装を行った。結果を図 4 に示す。図 4(a)が FPGA リソース使用量を、(b)が可視化された性能見積結果をそれぞれ示している。図 4(a)から、AI 推論モデルを ZCU102 ボード搭載の FPGA デバイスに実装する場合、ごく僅かな回路リソースと約 10%のメモリリソースのみが使用されることが分かる。この FPGA リソース使用量と推論性能（スループット）が実装方式によってどのように変化するかを調査した。FPGA リソース使用量については、

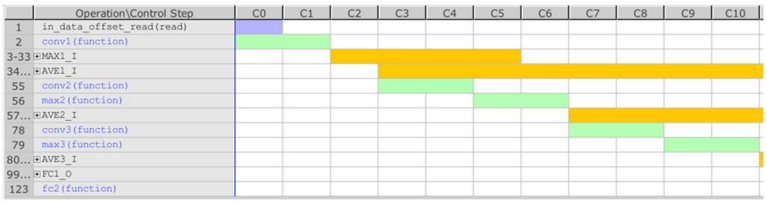


(a) 触覚センシングデータ (再生紙の例)　　(b) 紙識別における学習/検証時の精度

図 3　NN モデルの紙識別精度

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	926	-
FIFO	-	-	-	-	-
Instance	61	25	4312	5612	-
Memory	130	-	0	0	-
Multiplexer	-	-	-	923	-
Register	-	-	894	-	-
Total	191	25	5206	7461	0
Available	1824	2520	548160	274080	0
Utilization (%)	10	~0	~0	2	0

(a) FPGAリソース使用量



(b) FPGA性能見積結果

図 4　FPGA への推論モデル実装結果

FF / LUT に代表される論理回路リソースと BRAM メモリリソース、算術計算に特化した DSP リソースがあるため、それぞれの使用量の変化を調査・分析する。並列化もパイプライン化も実施しない素の回路を“オリジナル”とし、回路の並列化のみを実施したものを“並列化”、回路の並列化とパイプライン化の両方を実施したものを“並列化＋パイプライン化”と定義した。並列化とパイプライン化の概要を図 5 に示す。

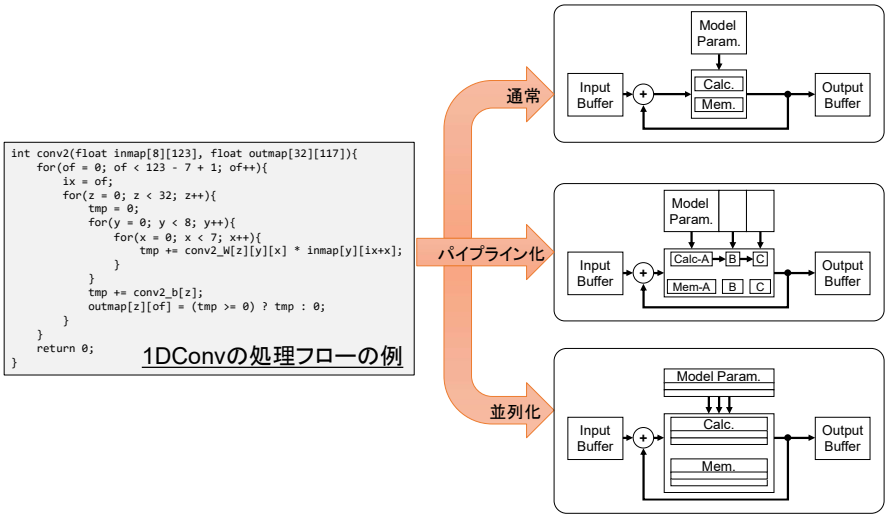


図 5　FPGA 実装時の並列化とパイプライン化

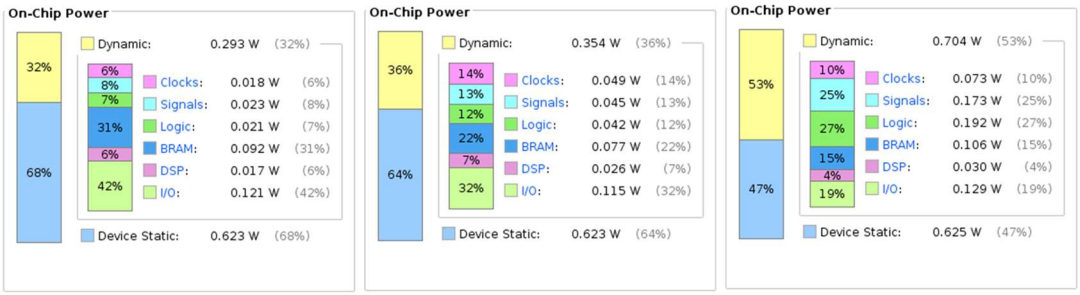
図は NN モデルの畳み込み層の処理を回路実装する例を示している。パイプライン化では層内の処理を演算単位ごとに細分化して別回路化することで、ループ内処理の途中で次の命令実行開始を可能とするものである。パイプライン化の説明では良く洗濯機の例が用いられるが、洗濯、すすぎ、脱水を別々のリソースで実現する場合、1 回目の脱水と同時に 2 回目のすすぎを、また同時に 3 回目の洗濯を行うことができ、1 回の洗濯の投入から脱水の完了までの遅延時間（レイテンシ）はパイプライン化の前後で不変だが、多数の洗濯を考える場合は処理性能（スループット）がパイプライン化によって向上する。同様に、並列化では NN モデルの層内の処理を複数の回路で同時並列に実行することで、処理性能（スループット）を並列化の前後で向上させることができる。それぞれのケースにおける FPGA リソース使用量と推論性能の結果を表 1 にまとめる。

表 1　FPGA への推論モデル実装結果

FPGA 実装	FPGA リソース使用量				推論性能
	BRAM	DSP	FF	LUT	[inf./sec]
オリジナル	191	25	5206	7461	0.91
並列化	198	59	17869	15980	2.94
並列化＋パイプライン化	198	51	25848	113344	4.17

表において、“並列化”では図 2 の NN モデルの層あたりの計算並列度を約 8 とした。一方で“パイプライン化”では層あたりのパイプライン段数を約 30 としている。FPGA 実装方式を“オリジナル”から“並列化”、”並列化＋パイプライン化”に変更するにつれて FPGA リソース使用量が増加する一方で推論性能が向上していく結果を得られた。

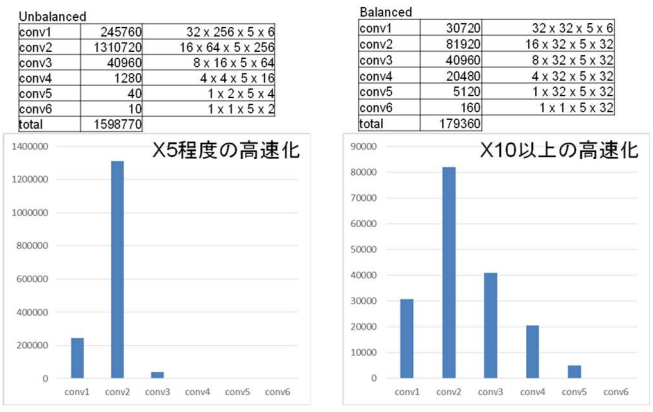
また、FPGA 実装結果の消費電力見積結果について図 6 に示す。



(a) オリジナル　(b) 並列化　(c) 並列化＋パイプライン化

図 6　消費電力の見積結果

図から FPGA リソース量に応じて FPGA の動作時消費電力 (Dynamic Power) が増加することが分かる。最後に NN モデルの構成と回路実装時の並列パイプライン化による高速化について調査を行った。CNN6 層モデルでの性能評価結果を図 7 に示す。図中の表とグラフは 6 つの畳み込み層のうち各層の積和演算回数を示しており、図 7(a)では、第 2 畳み込み層における積和演算回数が最大となる層構成になっている。このときの実装時最適化 (並列パイプライン化) による高速化の係数は約 5 が上限となった。これは各層の積和演算回数がアンバランスであり、パイプライン化や並列化によって計算資源を増やしても、メモリからのデータ読み出し性能が上がらず、処理性能のボトルネックとなってしまうためと考えている。一方で図 7(b)は枝刈りやモデル最適化によって、NN モデルの推論精度を維持したまま軽量化した場合の結果を示している。図では軽量化された NN モデルでは各層の積和演算回数がある程度平均化されており、回路実装時の並列パイプライン化によって処理性能を 10 倍以上に向上させることができた。この結果から、枝刈と並列パイプライン化との親和性は高いといえる。



(a) オリジナルNNモデル　(b) 軽量化後のNNモデル

図 7　NN モデル各層の演算回数と実装形態による高速化の例 (CNN6 層モデル)

＜引用文献＞

[1] 宮田博司他“エッジセントリックデジタルツインのためのエッジ AI&分散 DB コンピューティングアーキテクチャの一考察, ”VLD 研究会　VLD2023-111（2023.2）

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　1件／うち国際学会　3件）

１．発表者名 Yuji Yano, Kazutami Arimoto et. al.
２．発表標題 28-m W Fully Embedded AI Techniques with On-site Learning for Low-Power Handy Tactile Sensing System
３．学会等名 2022 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)（国際学会）
４．発表年 2022年

１．発表者名 有本和民
２．発表標題 おかやま組込みシステム・AI講座の取組について
３．学会等名 ソフトウェア信頼性研究会 第16回ワークショップ（FORCE2022（招待講演）
４．発表年 2022年

１．発表者名 杉野貴美廣、佐藤洋一郎、有本和民
２．発表標題 二輪車安全度向上のためのコーナー挙動測定系と評価について
３．学会等名 通信学会研究会（SWIM）
４．発表年 2022年

１．発表者名 杉野貴美廣、佐藤洋一郎、有本和民
２．発表標題 二輪車安全度向上のためのコーナーでの車両とライダーの挙動評価
３．学会等名 通信学会研究会（SWIM）
４．発表年 2022年

１．発表者名
久保智哉、横川智教、穂苅真樹、有本和民

２．発表標題
ドライバーの眠気予測を目的とした顔表情評定AIシステムの軽量化

３．学会等名
通信学会研究会（EMM）

４．発表年
2022年

１．発表者名
杉野貴美廣、佐藤洋一郎、有本和民

２．発表標題
ライダモニタリングによる二輪車安全向上システムの構想

３．学会等名
IEICE SWIM研究会

４．発表年
2021年

１．発表者名
有本和民、中村　宏、坂本龍一、鈴木悠太、武部秀治、吉川憲昭、木下研作、大矢智之

２．発表標題
エネルギーマッチングAIを用いるノーマリーオフ型ローカル5G基地局の概念実証

３．学会等名
IEICE RCS研究会

４．発表年
2021年

１．発表者名
H. Miyata, and K. Arimoto

２．発表標題
Localized data transfer system by Vehicles and Cellular Base Station in Local Area

３．学会等名
IEEE ICC’21 Workshop（国際学会）

４．発表年
2021年

１．発表者名 T. Kobayashi, T. Yokogawa, N. Igawa, Y. Sato, S. Fujii, and K. Arimoto
２．発表標題 A Compact Low Power AI Module Mounted on Drone for Plant Monitor System
３．学会等名 7th International Conference on Smart Computing and Artificial Intelligence (SCAI 2019), 2019.（国際学会）
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