FPGAによるAIアクセラレーションに適したニューラルネットワークモデル評価

伊田　直也　　増渕　悠二

近年、さまざまな分野でAIの応用が進み、大規模なサーバーサイドでのAIだけでなく、軽量なモデルを小型のエッジデバイス上で推論するシステムも積極的に開発されている。このエッジAIを実現する一つのアプローチとして、ARMプロセッサーとFPGA（Field Programmable Gate Array）を搭載したMPSoCと呼ばれるチップをエッジデバイスに搭載するアイディアがある。これはFPGAでAI推論をアクセラレーションすることで、推論速度の飛躍的な改善を目指している。

OKIアイディエスはソフトウェア設計やFPGAの開発経験が豊富にあり、FPGAを利用したAIアクセラレーションにも取り組んでいる。しかし、AIモデルは非常に多岐にわたるため、カスタムモデルがどの程度アクセラレーションできるかという見積もりは困難である。そのため、テストモデルをまず実装し、推論速度を評価するアプローチでの開発プロセスをとっている。この課題を解決するために、FPGAアクセラレーションと親和性が高いニューラルネットワーク（NN）モデルを明らかにするという目的で検討した。その結果、大規模なモデルでFPGAアクセラレーションはより効果的との結論を得た。本稿では、この検討結果について紹介する。

FPGA を利用した AI アクセラレーション

FPGAはプログラミングが可能な集積回路であり、並列計算を得意としたデバイスである。適切に回路情報（IPコア）を設計することで、特定用途ではCPUよりも高い演算性能を出すことができる。この特性を生かしてFPGAデバイスは画像処理やデジタル信号処理をはじめとする幅広い用途で使われ、近年ではAIアクセラレーターとしてNNの計算でも使われている。

FPGAを用いてAIアクセラレーションする場合には、NNモデルに合わせてIPコアを個別に開発することもできるが、開発ベンダが提供するFPGAの推論フレームワークを利用することが多い。これらの推論フレームワークの多くは、推論を実行するFPGAのIPコアとそのIPコアを利用するためのライブラリー、学習済NNモデルをFPGA推論に合わせて変換するツールなどから構成される。

AIアクセラレーション用のIPコアはさまざまなNNモデルに対応できるように汎用性を持たせて設計され、畳み込み層やプーリング層、全結合層などのNNレイヤーの基本機能をもつユニットを多数並列に実装している。また並列度を上げるために、多くの場合32bitのNNモデルを8bitに量子化して推論する。そのためFPGAを用いたアクセラレーションでは推論速度は大きく改善するが、認識精度は若干低下することが知られている。

一般的なFPGAを用いたAIアクセラレーションのフローを図1に示す。まずFPGAでAI推論を行うNNモデルを8bitに量子化する。これは学習済みNNモデルごとに一度だけ実行すればよい。続いて推論処理を開始する前に、量子化したNNモデルの重み情報をFPGA上に転送する。その後バッチと呼ばれるAI推論処理の単位ごとにデータが入力され、前処理、バス書込、FPGA処理、後処理、バス読込が繰り返されて推論が行われる。このFPGAを利用した推論の逐次処理がCPU単体での推論時間より早い場合には、AIアクセラレーターとして有効である。

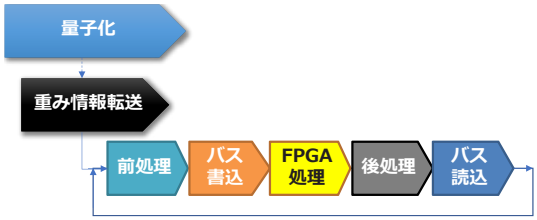


図 1　FPGA を用いた AI アクセラレーションのフロー

FPGA アクセラレーションに適した NN モデル

(1) 利用モデル

まず本検討では単純なCNN（Convolutional Neural Network:畳み込みNN）モデルとして広く知られるLeNet[2)]を基にして、手書き数字を認識するカスタムモデルを作成した。このNNモデルの入力は28x28ピクセルのMNISTデータベースとし、出力は0から9の数字での評価結果を返す。このNNモデルのレイヤー構成を図2に示す。

一般にCNNモデルでは推論時の計算量の大部分が、畳み込み層で占められる。そこでモデルの違いによるFPGAアクセラレーションの効果を調査するために、下記のアプローチで畳み込み層を拡張した。

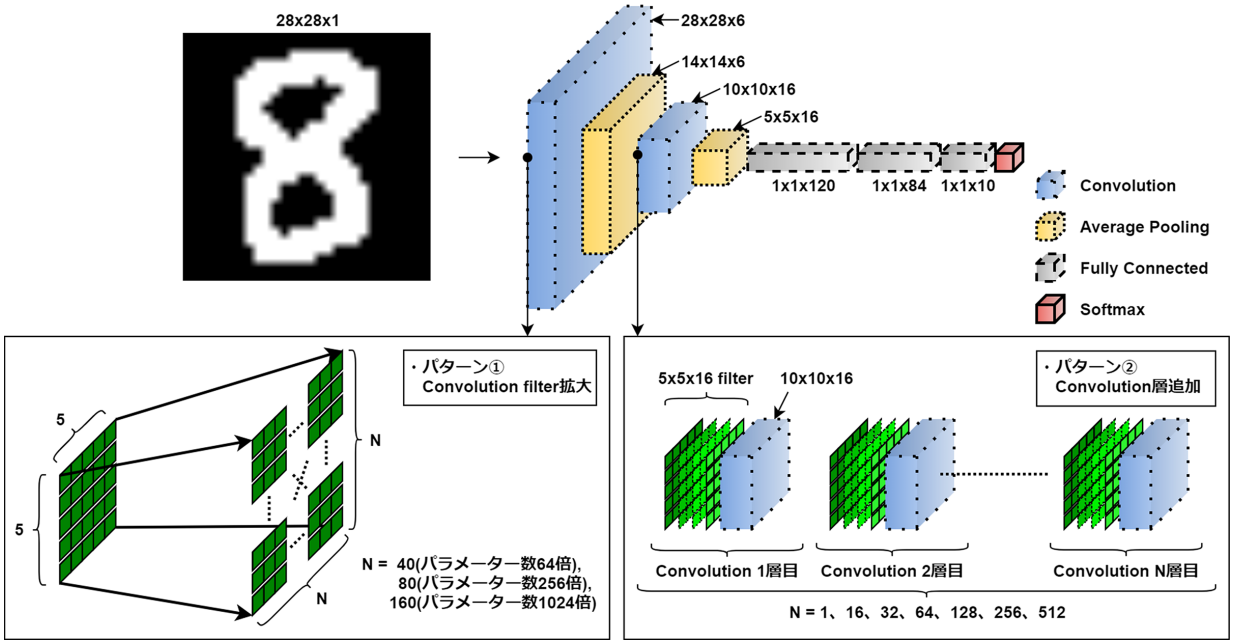


図 2　手書き文字認識 NN モデルのレイヤー構成

・パターン①　レイヤーサイズ拡大
・パターン②　レイヤー層数増加

いずれの拡張もパラメーター数の増加が目的である。パラメーター数とはNNの「重みの数」であり、推論時の計算量はパラメーター数に依存する。計算量の大小だけでなくNN構造の違いがAIアクセラレーションにどのような影響を与えるのか、この拡張により確認した。

具体的にはパターン①ではレイヤーサイズ拡大の影響を確認するために、入力層に近い畳み込み層のフィルターサイズを変えることでパラメーター数を増加した。(**図2** パターン①)フィルターサイズはオリジナルモデルを1とした場合に、64、256、1024倍のNNモデルとした。対してパターン②のレイヤー層数増加の影響確認では、NN内の畳み込みレイヤーの層数を増やすことでパラメーター数を増加した。(**図2** パターン②)拡張したレイヤーは16、32、64、128、256、512層である。パターン①及びパターン②で生成したNNモデルのパラメーター数を、**図3**と**図4**にそれぞれ示す。

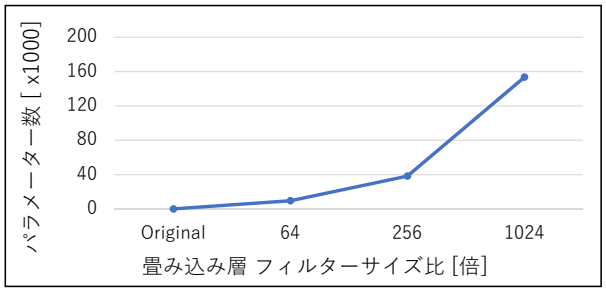


図 3 パターン①　レイヤーサイズ拡大とパラメーター数

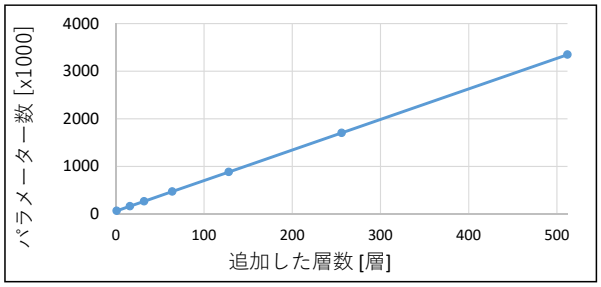


図 4 パターン②　レイヤー層数とパラメーター数

(2) 調査環境構築

本検討で使用した調査環境を**表1**に示す。

表 1　調査環境

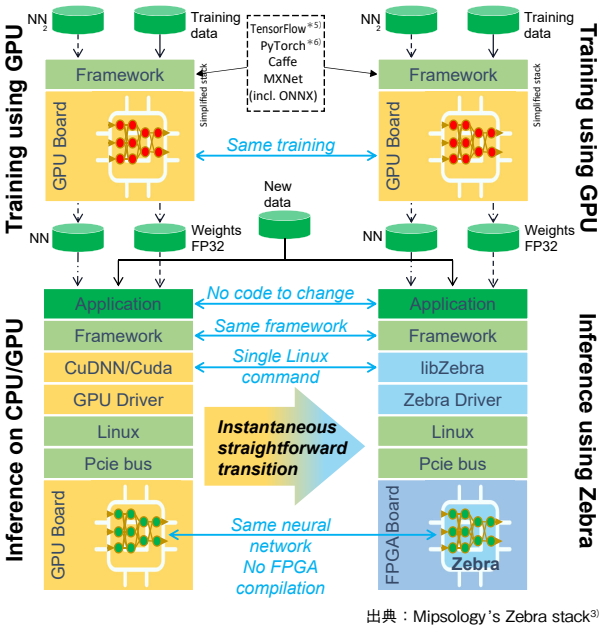
名称	種類
CPU	Intel® Xeon® *1) W-2123
メモリー	64GB (DDR4 2666)
FPGA カード	Xilinx® Alveo™ *2) U50
OS	Ubuntu*3) 18.04.03
AI アクセラレーション プラットフォーム	Mipsology Zebra V2021.5.3

FPGAはAMD社のAlveo U50データセンターアクセラレータカードを利用した。このカードはPCからのFPGA利用を目的としたデバイスであり、PCIeスロットを介してホストPCと接続する。本検討ではFPGAでのアクセラレーションの傾向を分析することを目的として、PCとFPGAカードの組合わせで調査した。

*1)Intel Xeonはアメリカ合衆国およびその他の国におけるIntel Corporationまたはその子会社の商標または登録商標です。　*2)XilinxおよびAlveoはAdvanced Micro Devices, Inc.の商標です。　*3)Ubuntuは、Canonical Ltd.の商標または登録商標です。

またAIアクセラレーションプラットフォームとしてMipsology[*4]社のZebra[*4]を利用した。Zebraプラットフォームは FPGAの回路情報と、その回路情報を利用するための各種ライブラリー、ドライバーから構成される。Zebraプラットフォームの概要を図5に示す。

ZebraプラットフォームはCPUやGPU向けに開発されたオリジナルのNNを再学習することなく自動で量子化し、そのまま推論に利用することができるのが、大きな特徴である。本検討でも、Zebraプラットフォーム利用のためのスクリプトの変更は一切必要なかった。



出典：Mipsology's Zebra stack[3]

図5　Zebra プラットフォーム概要

(3) パラメーター数の増加による影響

まずパターン①のレイヤーサイズ拡大が推論時間に影響することを確認するために、入力層付近での畳み込み層のフィルターサイズを変化させたNNモデルの推論時間をCPUとFPGAで比較した。結果を図6に示す。またFPGAの推論時間の内訳を図7に示す。

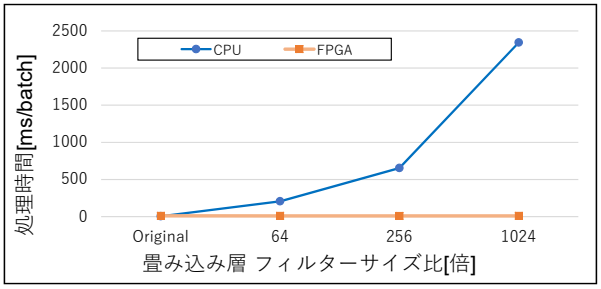


図6　レイヤーサイズ拡大による推論時間の変化

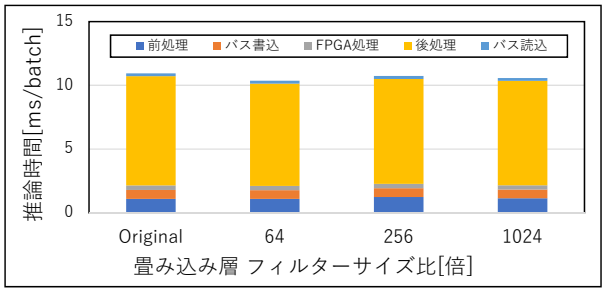


図7　レイヤーサイズ拡大での FPGA 推論時間内訳

図6より、フィルターサイズ（≒レイヤーサイズ）が大きくなった場合にはCPUの推論時間が大きく増加しているが、FPGAの推論時間にはほぼ変化がなかった。CPUの推論時間の変化の傾向は図3とほぼ一致し、CPUの推論時間はパラメーター数に依存していると判断できる。対してFPGAの場合には、レイヤーサイズが、推論時間には影響しない。FPGAの推論時間の内訳である図7を確認しても、レイヤーサイズの変化による影響は見られない。レイヤーサイズが1024倍（パラメーター数160,000）程度であれば、本環境のFPGAで実施した場合には、瞬時に計算されることを意味する。

続いてパターン②のレイヤー層数の増加が推論時間にどのような影響を与えるかを確認するために、畳み込み層を最大で512層追加し、CPUとFPGAで動作を比較した。結果を図8に、またFPGAの推論時間内訳を図9に示す。

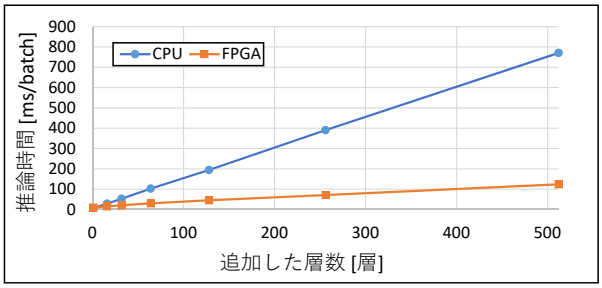


図8 レイヤー層数変化による推論時間の変化

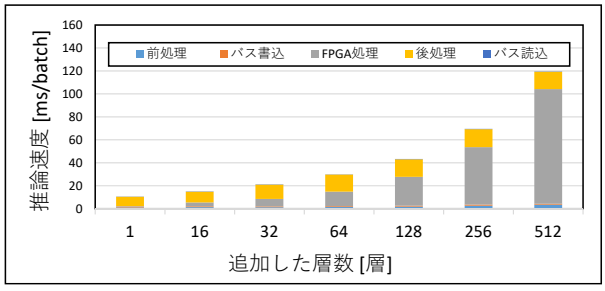


図9　追加したレイヤー層数ごとの FPGA 推論時間内訳

*4) Mipsology及びZebraはMipsology SASの登録商標です。　*5) TensorFlowはGoogle, Inc.の登録商標です。　*6) PyTorchは、Facebook, Inc.の登録商標です。

図8より、CPU·FPGAともにレイヤー層数と推論時間には比例関係があることが分かる。これは図4に示したモデルのパラメーター数と同じ関係であり、CPU及びFPGAの推論時間は、モデルのパラメーター数に依存することが確認できる。CPUとFPGAの推論時間を比較すると、レイヤー層数が少ない領域では差は少ないが、レイヤー層数が増えるに従い推論時間の差が大きくなり、512層ではFPGAの推論時間がCPUの6分の1になった。図9に示したFPGA推論時間内訳を確認すると、層数が増えるに従いFPGA処理のプロセスが推論時間に占める割合は大きくなる。対してFPGA処理以外のプロセスであるバスの読書き、前処理や後処理の時間は、層数による影響はほとんどない。このことから、レイヤー層数の多いモデルの方が、推論時間内に占めるFPGA処理の割合が増えるため、FPGAでの推論により適したモデルであると言える。

(4)FPGAアクセラレーションに適したモデルの分析

CPUとFPGAの推論時間の比較結果より、FPGAでアクセラレーションを行うのに適しているのは、「レイヤーサイズが大きく、レイヤー層数も多い大規模なCNNモデル」と判断できる。これはCPUに比べてFPGAは動作クロックが低いが、リソースが許す限り並列計算できるために、パラメーター数の大きいモデルでは顕著にCPUよりも推論時間が短くなることを意味する。当然ながらFPGAで同時に行える並列計算には上限があり、FPGA上にプログラミングされたIPコアの内容に依存する。大規模なNNモデルをFPGA上でアクセラレーションするためには、モデルサイズに見合ったIPコアとFPGAデバイスの選定が必要になる。

また同じパラメーター数の増加でも、パターン①のレイヤーサイズ拡大の方が、FPGAアクセラレーションとの相性がよいことが確認できた。このパターン①とパターン②のパラメーター数による推論時間の変化傾向の差は、次のように説明できる。一般にNNモデルでは、あるレイヤーの出力が次のレイヤーの入力となり、入力層から出力層まで各レイヤーの処理を行いながら情報が伝搬することで推論する。パターン①で要求される同一レイヤー内の演算は、FPGAでは一度に並列演算することができるため、レイヤーサイズ拡大によるパラメーター数の増加は推論時間への影響が少ない。対してパターン②のようにレイヤー層数が多い場合には、レイヤー間の伝搬の回数が多いため、計算処理にレイヤー層数分の順序性が必要になる。この違いが、パターン①とパターン②でFPGAアクセラレーションの傾向の違いとして現れたと考えられる。

まとめ

本稿ではFPGAによるAIアクセラレーションに適したNNモデルを検討し、大規模なCNNモデルではより効果が期待できることを報告した。特にレイヤーサイズを拡大するNNモデルの変更は、FPGAアクセラレーションの効果を最大限に得ることができる。OKIアイディエスではこの知見を活用し、FPGA化を見越したNNモデルのカスタマイズとAIアクセラレーションを組み合わせたソリューションの提供を開始する予定である。 ◆◆

■参考文献

1)伊田直也:FPGAとPCASによるモデル軽量化を利用したAIアクセラレーション、OKIテクニカルレビュー第238号、Vol.88 No.2、pp.62-65、2021年11月
2)Yann LeCun, Patrick Haffner, Léeon Bottou, Yoshua Bengio, Object Recognition with Gradient-Based, Shape, Contour and Grouping in Computer Vision, 1999.
3)Deep Learning Inferencing with Mipsology using Xilinx ALVEO on Dell EMC Infrastructure、2022年3月4日、https://downloads.dell.com/manuals/common/deeplearning_mipsology_poweredge.pdf

●筆者紹介

伊田直也:Naoya Ida. 株式会社OKIアイディエス 事業統括部 開発部
増渕悠二:Yuji Masubuchi. 株式会社OKIアイディエス 事業統括部 開発部

TIPS【基本用語解説】

MPSoC(MultiProcessor System on a Chip)
AMD社が提供する、複数のARMプロセッサー、FPGA、メモリー、I/Oなどを搭載したチップ。

CNN(Convolutional Neural Network)
畳み込み層を持つニューラルネットワークの総称で、機械学習の中でも画像認識の分野で広く使われるニューラルネットワークである。畳み込み層、プーリング層、全結合層から構成される。

MNISTデータベース
28x28ピクセルの0～9の手書き数字画像のデータベース。60000枚の訓練用画像と10000枚の評価用画像が含まれている。