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

伊田　直也

近年ではAIの進歩は目覚ましく、非常に多岐にわたる領域で活用が進み、人々の生活を便利にしてくれている。またニューラルネットワーク（NN）モデルも、日々新しいモデルが研究・開発されている。ここでエッジデバイスでのAIの実現性に目を向けてみると、高速化や省スペース、省電力などのさまざまな課題がある。

これらの課題は「NNの軽量化」と「ハードウェアアクセラレーション」を用いることで、大きく改善が期待される。NNの軽量化はより少ないリソースでのAI推論を可能にし、ハードウェアアクセラレーションは高速化だけでなく省エネルギーの観点でも有効である。

本稿では、OKIアイディエスで実現したFPGAデバイスによるアクセラレーションとモデル軽量化技術PCASを組み合わせて、エッジ装置でも利用可能なAIアクセラレーションを紹介する。

FPGAを利用したAIアクセラレーション

FPGA（Field Programmable Gate Array）はプログラミングが可能な集積回路であり、並列計算などを得意としたデバイスである。適切に内部回路を設計することでCPUやマイクロプロセッサーに比べて、特定用途ではより高い電力当たりの計算性能を出すことができる。そのため画像処理や暗号化、通信などの幅広い用途で使われ、近年では高速フーリエ変換やNNの計算にも使われている。

一般にFPGAの内部回路開発は、Verilog HDLやVHDLなどのHDL（Hardware Description Language:ハードウェア記述言語）を利用して論理回路を設計し、論理合成ツールを介して回路情報を生成することで行う。FPGAの開発は、動作タイミングを考慮した設計やハードウェアの知識が必要になるため、FPGA専門の技術者が必要とされる場合が多い。

FPGAの開発経験が豊富なOKIアイディエスでは、FPGAを利用したAI推論に対して以下のアプローチで取り組んできた。

A）NNに合わせて回路情報を開発する
B）開発ベンダ提供のAI開発ツールを利用する
C）FPGA AIアクセラレーションプラットフォームを利用する

NN高速化の最初のアプローチA）として、公開されているNNのC言語ソースプログラムを用いて、その畳み込み層と全結合層を高位合成ツールによりFPGA化した。苦労してFPGA化できたものの、ソフトウェアの流れをそのままFPGA化したため、期待したほど性能向上はできなかった。内部演算を固定小数点化する際の量子化誤差低減、及び全体フローの最適化が課題として残った。この課題の解決にはコストとスケジュールの両面で明らかにリスクがあり、このアプローチの継続には慎重な判断が必要であった。

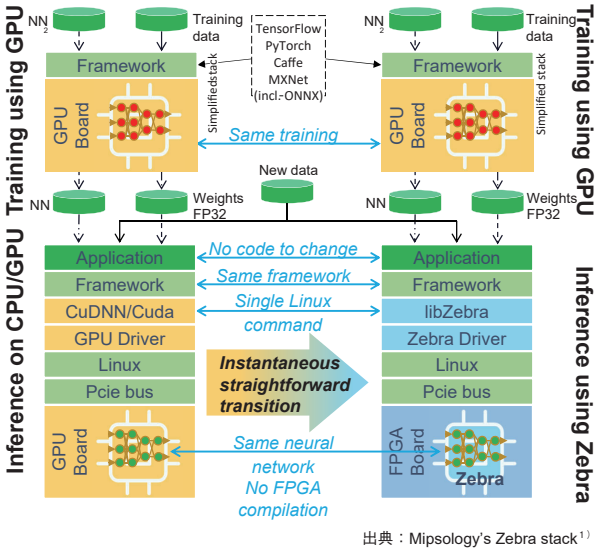
その頃から開発ベンダのAI開発ツールが急速に利用できるようになり、実際にOKIアイディエスでもMPSoCなどのエッジデバイスやFPGA搭載のPCIeカードへのNNインプリメントを行うようになった。このアプローチB）は、A）に比べて開発期間を大幅に短縮できるようになった。本アプローチの結果、当社採用のNNモデルでは、CPUと比較してFPGAによる推論では数十倍～数百倍の速度向上が見られた。公開されている著明なNNモデルの場合には本アプローチによって期待される高速化効果を得られたが、実装置上で求められるAIは非常に多岐にわたるユーザーカスタムNNモデルのインプリメントが要求され、AI開発ツールが対応できているのはその一部に留まるのが実情である。そのためカスタムNNモデルをFPGA上で動作させるには、多くの時間を要する。

そこで次なるアプローチC）として、OKIアイディエスではMipsology*1)社のFPGA AIアクセラレーションプラットフォームZebra*1)を利用したNN高速化に取り組むこととした。

Zebraプラットフォームは FPGAの回路情報と、その回路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。図1にZebraプラットフォームを利用したアクセラレーションの概略を示す。Zebraプラットフォームは、GPUやCPU向けに開発されたオリジナルのNNを、簡単にFPGA上でアクセラレーションできるのが最大の特徴である。Zebraライブラリーはアプリケーションから TensorFlow*2)やPyTorch*3)などの一般的なAIフレームワークを介して利用できるように設計され、CPUやGPU向けに行われたトレーニング済みNNを再学習する必要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。その他の特徴とし

*1）Mipsology及びZebraはMipsology SASの登録商標です。　*2）TensorFlowはGoogle, Inc.の登録商標です。　*3）PyTorchは、Facebook, Inc.の登録商標です。

ては、Zebraプラットフォームは高スループット、低レイテンシーであり、またデータセンター向けの大規模コンピューティングから、エッジや組込み領域まで、あらゆるサイズのFPGA上で利用できる。



出典：Mipsology's Zebra stack[1)]

図1　Zebra プラットフォーム概要

実際にはアプローチA）及びB）では、FPGAを利用する環境に合わせたAI推論スクリプトの書換えや、ライブラリーやドライバーの実装など、FPGAだけでなく関連ソフトウェアの開発にも時間を要していた。Zebraプラットフォームを利用するアプローチC）では、ドライバーやライブラリーがAI推論スクリプトの変更なしで実行できるようにパッケージ化され、この点でも開発時間を大幅に短縮することができた。

AIモデル軽量化技術PCAS

エッジデバイスでのAI推論の高速化には、ハードウェアアクセラレーションだけでなく、NNそのものの軽量化も重要である。軽量なモデルはより少ないリソースでのAI推論が期待でき、また演算量が減るために推論速度にもポジティブな影響を与えることが一般的に知られる。

PCAS（Pruning Channels with Attention Statistics）は、OKIが開発したモデル軽量化技術であり、CNN（Convolutional neural network）モデルを対象としたチャネル単位の枝刈りを最適に行う手法である。軽量化するCNNモデルの層間にアテンションモジュールと呼ばれる新たなNNモデルを追加する。このモジュールを利用して学習するなかで全ての層の全てのチャネルの重要度を評価し、重要度の低いチャネルから削減していくことで、精度を維持しながらメモリー使用量と演算量を削減する。人手による層ごとのチャネル削減率の設定が不要であり、ネットワーク全体でのチャネル削減率を設定するだけで、重要度の低いチャネルから削減が自動で行われることがPCASの最大の特徴である。

PCASとFPGAを利用したアクセラレーション

本章では、Zebraプラットフォーム環境を構築し、PCASにより軽量化したモデルによる推論をFPGAでアクセラレーションした事例を紹介する。

(1) 実行環境構築

本開発で使用した開発環境を表1に示す。

表1　開発環境

名称	種類
CPU	Intel Xeon W-2123
メモリー	64GB (DDR4 2666)
FPGAボード	Xilinx Alveo U200
OS	Ubuntu 18.04.03
Zebraプラットフォーム	V2021.1.3

FPGAはXilinx社のAlveo U200 データセンターアクセラレータカードを採用した。Alveo U200はPCからのFPGA利用を目的としたデバイスであり、PCIeスロットでホストPCと接続できる。

FPGAへのZebraコア（Zebraプラットフォームで利用するFPGAの回路情報）の書込みは、Xilinx社により提供されているXRT（Xilinx RunTime Library）コマンドを利用して行った。推論スクリプトの実行は、Linux OS（Ubuntu 18.04）上のZebraプラットフォームをインストールしたDocker環境内で行った。CPUで実行可能なスクリプトとNNを作成後、FPGAで推論を行うまでに要した時間は5時間程度であった。

(2) 利用モデル

本検討で利用したモデルは、PyTorchで実装したVGG11であり、評価に使用したデータセットはCIFAR-10である。PCASを利用してモデルを軽量化し、10段階に分けてチャネル削減量の異なるファイル（#1～#10）を作成した。未圧縮のファイルサイズが37.64MB、最も軽量化したファイルが0.52MBである。作成した各ファイルのサイズを図2に示す。なお、#0が未圧縮のオリジナルファイルで#10が最も軽量化したファイルである。

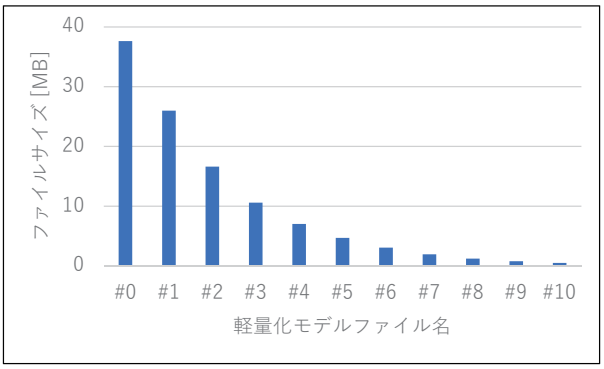


図 2　各モデルのファイルサイズ

また各モデルファイルのパラメーター数と精度の関係を示したのが、図3である。パラメーター数は演算量に直結する。軽量化を進めることでチャネル間のパラメーターが減るため、推論に必要な演算量が減る関係がある。なお、精度とはCIFAR-10のデータの認識率である。

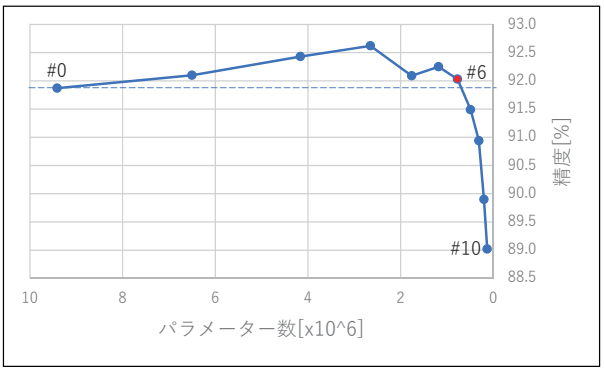


図 3　精度とパラメーター数の関係

左上がオリジナルモデルの#0であり、右に向かうに従い、チャネル削減率が高くなる。オリジナルのモデルの認識精度は91.87%であり、軽量化モデルファイル#6までの軽量化であれば、精度を保ってチャネル削減できていると判断できる。このモデルを利用して、PCASによる軽量化の前後、及びZebraプラットフォーム利用の前後での推論時間・精度を比較した。

(3) 実行結果

まず、Zebraプラットフォーム及び、PCASの効果を確認するため、各モデルファイル画像1枚あたりの推論時間を比較したものを、図4に示す。

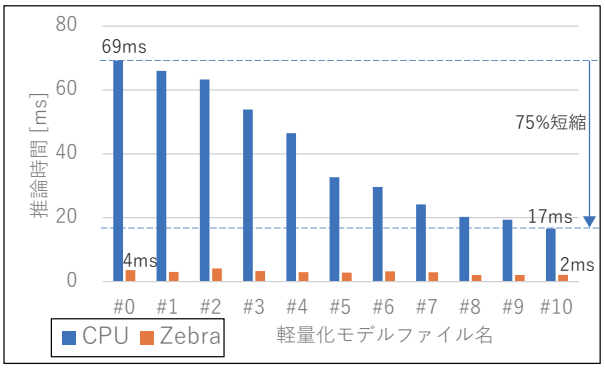


図 4　画像 1 枚あたりの平均推論時間

いずれのモデルでもCPUでの推論に比べてZebraプラットフォームを利用した推論では、処理時間が大幅に短縮され、FPGAによるアクセラレーションの効果を確認できる。未圧縮のモデルファイル#0を例にとると、CPUでの推論時間69msに比べて、Zebraプラットフォームを利用した場合は4msであり、約5.8%に短縮できた。

PCASの軽量化効果は、CPUの推論時間に着目すると顕著である。未圧縮ファイル#0から最も軽量化したモデルファイル#10の間で推論時間を75%短縮している。Zebraプラットフォームを介してFPGAを利用して推論した場合にも、未圧縮の場合には約4ms程度要していた推論時間が、軽量化後には約2msになり、半分近い推論時間の短縮を確認できた。

次に、CPUとZebraプラットフォーム間の推論精度に差がないことを確認するために、各モデルファイルでの推論精度で比較したものを図5に示す。

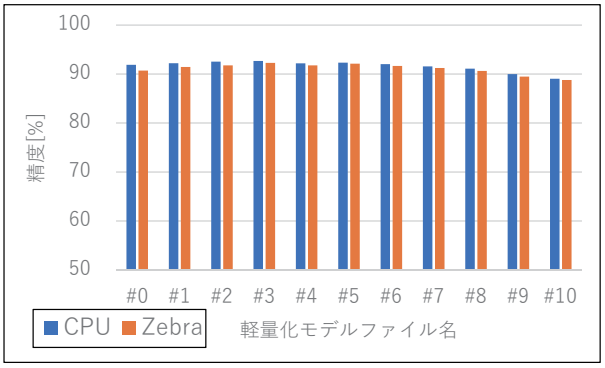


図 5　平均推論精度

Zebraプラットフォームによる推論の精度は、CPU推論と比較してほぼ同等の水準にあるが、若干の精度低下が生じた。一般にFPGA上でAI推論を行うためには、回路規

模を抑えつつ演算速度を確保するために、重みパラメーターのビット数を削減する量子化を行う必要があり、その際に丸めの誤差による性能低下が生じる。Zebraプラットフォームを利用した場合にも、この精度低下が生じた。

軽量化の程度と精度の落込みに着目すると、未圧縮#0の場合が1.19%と最も落込みが大きく、軽量化が進むにつれて精度低下が少なくなる傾向が見られた。#10では、落込みは0.25%にまで減少する。この点から、Zebraプラットフォームによる量子化と、PCASによるモデル軽量化の親和性が高いことを確認できる。

(4) FPGAとPCASを組み合わせたアクセラレーション

軽量化前のモデル(#0)によるCPU推論での精度91.87%と比較すると、#6のモデルまではZebraプラットフォームを利用した推論でも、軽量化前とほぼ同水準の精度が得られた。このモデルはPCASによりオリジナルモデルに対して約8%のサイズに軽量化したモデルである。PCASとZebraプラットフォームを組み合わせて利用することで、約8%に軽量化したモデルでも、オリジナルモデルから精度を落とさずにアクセラレーションすることができた。

今後の展望

本稿ではFPGAを利用したAIアクセラレーションとモデル軽量化を組み合わせて、より省リソースかつ高速なAI推論を実現するOKIアイディエスの取組みを紹介した。

現在は、Zynq*4) UltraScale+ MPSoCなどのFPGAを搭載したエッジデバイス上でのAI推論システムを、PCASとZebraプラットフォーム上に実装を進めている。さらにFPGA内の別領域にAI推論以外の機能(画像処理、高速通信など)も合わせて実装することで、より付加価値の高いAIエッジデバイスの実現が目標である。またZebraプラットフォームとPCASを組み合わせたAI推論ソリューションを、ユーザーに提供開始する予定である。 ◆◆

■参考文献

1) Deep Learning Inferencing with Mipsology using Xilinx ALVEO on Dell EMC Infrastructure、2021年9月1日、https://downloads.dell.com/manuals/common/deeplearning_mipsology_poweredge.pdf
2) 山本 康平、橘 素子、前野 蔵人:ディープラーニングのモデル軽量化技術、OKIテクニカルレビュー第231号、Vol.87 pp.24-27、2019年5月

*4) ZynqはXilinx, Inc.の登録商標です。

●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

TIPS【基本用語解説】

VGG11
VGGは畳み込み層と全結合層で構成される比較的単純なNNであり、画像認識の分野で広く使われている。レイヤー数によってバリエーションがあり、VGG11は11層からなるNNである。

CIFAR(Canadian Institute For Advanced Research)-10
画像認識評価で広く使用されるデータセット。32x32のRGB画像60000枚で構成され、各画像は10種類のいずれかのクラスに分類されている。