Vol.2016-ARC-222 No.12
2016/10/6

ディープラーニング向けアクセラレータアーキテクチャのFPGA実装

林 遼[1]　森下 真幸[1]　高田 遼[1]　坂本 龍一[1]　近藤 正章[1]　中村 宏[1]

概要：近年，高電力効率を目的としたディープラーニングアクセラレータの研究が活発に行われている．しかし，データアクセスやネットワーク構造を工夫することで省電力化を行う研究が多く，対象とするネットワーク構成が限定され拡張性に乏しいという課題が考えられる．そこで我々は，命令セットにより多様なネットワークに柔軟に対応可能なアクセラレータ構成を検討している．今回，アーキテクチャのプロトタイプを FPGA を用いて実装し，畳み込みニューラルネットワークによる画像認識を実行するデモンストレーションを製作した．また，いくつかのパラメータを変化させ畳み込みニューラルネットワークの実行時間について見積りを行った．

1. はじめに

近年，自動車やモバイル機器などの組み込みシステムでディープラーニング技術を利用することが期待されている．そこで，限られた電力資源下でも動作する高電力効率なディープラーニング専用アクセラレータの開発は重要な課題となっている．

組み込みシステム向けの低電力アクセラレータとしては，Chen らの Eyeriss[1] が知られている．Eyeriss では消費電力を抑えるために，チップ内バッファや演算ユニット間ネットワークなどを工夫し，外部メモリアクセスの削減に成功している．また Reagen らは，消費電力を抑えながらもディープラーニングの推定精度を落とさずに実行できるアクセラレータのための設計手法 Minerva[2] を提案している．

これらの従来研究では，特定のネットワーク構成向けに最適化されたアクセラレータに関する研究や，データアクセス削減のためにネットワーク構造を縮小させるものが多い．対象とするネットワーク構成が限られ，様々なネットワークへの拡張が簡単ではないという課題がある．それに対して，我々は多様な種類のネットワーク構成に対応可能なプログラマブルなディープラーニング向けアクセラレータを検討している．例えば，画像認識に用いられる畳み込みニューラルネットワーク (Convolutional NeuralNetwork: CNN) は，畳み込み層，プーリング層，全結合層などが多層に積み重なったネットワーク構成をとるが，本研究のアクセラレータは多様な構成を持つ CNN を実行可能である．

本稿では，アクセラレータ構成検討のために，ARM 混載 FPGA ボードへ実装した 7 層 CNN による画像認識アプリケーションを実行するシステムについて述べる．また，アクセラレータの性能を評価するために，いくつかのパラメータを変化させ 7 層 CNN の実行時間について見積りを行った．

2. アクセラレータアーキテクチャ

2.1 全体構成

本研究で検討しているアクセラレータは，コアを複数持つマルチコア構成をとる．4 コア構成のブロック図を**図 1** に示す．各コアは，SIMD 型演算器と制御ユニット，命令メモリ (inst)，ストリームバッファ (sbuf)，データメモリ (dmem)，ルックアップテーブル (lut)，データ出力用メモリ (omem) の 5 つのメモリを持つ．

sbuf は再利用性の低いデータに，dmem は再利用性の高いデータに用いる．subf はダブルバッファリングを行うことで，DMA により外部メモリと sbuf 間でデータ転送中に演算処理をオーバーラップさせ，遅延を隠蔽可能である．また，CNN を始めとした多層のニューラルネットワークを複数コアで実行する際，演算結果をコア間で共有する必要があるため，出力用メモリの omem は各コアで共有されている．

2.2 コアのアーキテクチャ

命令セットは 16 ビット固定長である．現在の実装では命令パイプライン化が十分ではないが，今後パイプライン化された実装に拡張する予定である．ディープラーニング

[1] 東京大学

© 2016 Information Processing Society of Japan

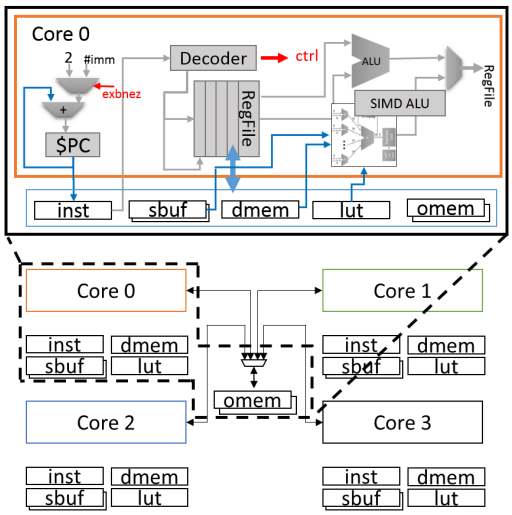

図 1　4 コア構成のアクセラレータ

専用アクセラレータでは，膨大な積和演算を効率よく実行できることが重要であるため，各コアは SIMD 型積和演算器と独自の SIMD 算術命令を持つ．処理対象データは sbuf と dmem 格納されており，この SIMD 型積和演算器へダイレクトに供給される．今回の実装では，SIMD 型積和演算器は 16bit x 4，あるいは 8bit x 8 で演算を行い，sbuf と dmem のデータバスはそれぞれ 64 ビット幅となっている．また，SIMD 算術命令は CNN で必要な畳み込み演算のアクセスを効率良く行うことができるよう設計されている．コアの基本構成要素としてはこれらの他に，32 ビットのレジスタファイル 16 本と制御用マイクロコントローラがある．

3. FPGA 評価ボードを用いた実装

開発したアクセラレータのデモンストレーションとして，7 層の CNN による画像認識アプリケーションを実行するシステムを FPGA ボードに実装した．システムのデータフローは図 **2** のようになっている．使用した FPGA ボードは Zynq-7000 SoC ZC702 で，4 コア構成のアクセラレータをプログラマブルロジック部に実装した．また，ホスト CPU には Zynq に搭載されている ARM Cortex-A9 CPU を使用した．アクセラレータの各メモリにはデュアルポートのブロック RAM を用いた．ホスト CPU とアクセラレータ部の動作周波数はそれぞれ 667MHz と 25MHz である．

4. 性能見積り

今回 FPGA ボードを用いて実装した結果，アクセラレータ部の動作周波数は 25MHz となったが，現在アクセラレータチップを LSI 上に試作中である．そこで本稿では，パラメータを動作周波数とした場合の実行時間の見積りについて述べる．評価アプリケーションには 7 層の CNN を用いた．また，今回の見積りではホスト CPU の動作周波数を 667MHz に固定したため，ある動作周波数以上になるとホスト CPU の主記憶とブロック RAM 間のデータ転送コストで実行時間が律速する．見積りの結果，実行時間は動作周波数に対して図 **3** のように変化し，アクセラレータの演算性能と外部メモリとのデータ転送速度の均衡がとれる動作周波数は 50MHz となった．

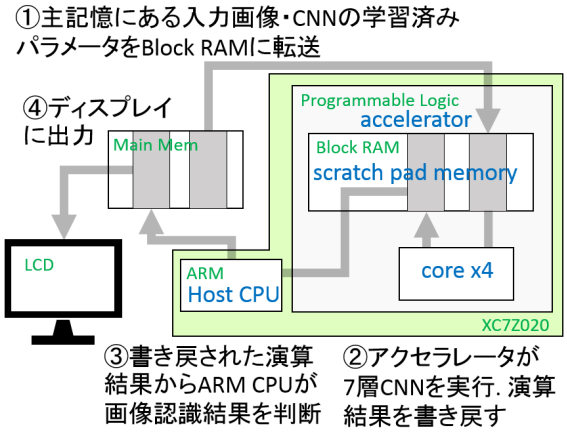

図 2　デモシステムのデータフロー

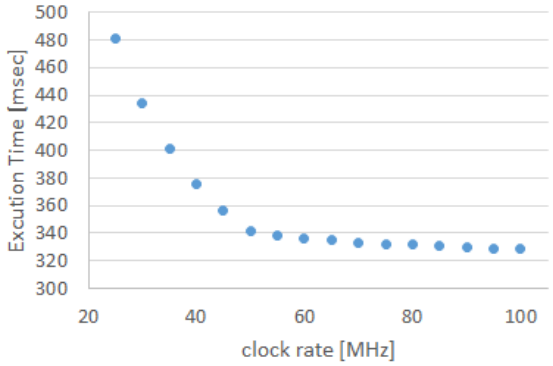

図 3　7 層 CNN の実行時間見積り

5. おわりに

本稿では，ディープラーニング向けアクセラレータの検討のため，4 コア構成のアクセラレータを FPGA ボードに実装し，畳み込みニューラルネットワークによる画像認識アプリケーションのデモシステム開発について述べた．また，アクセラレータの動作周波数パラメータとして実行時間を見積り，性能を評価した．

謝辞　本研究の一部は JSPS 科研費基盤研究（S）25220002 の助成によるものである．

参考文献

[1] Chen, Yu-Hsin, et al. ”Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks.” 2016 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2016.

[2] Reagen, Brandon, et al. ”Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators.” Proceedings of the 43rd International Symposium on Computer Architecture, ISCA. 2016.

© 2016 Information Processing Society of Japan