7J-02

# FPGA 上での CNN パラメータ動的更新手法の性能評価 *

青戸武蔵 †1　和田康孝 ‡1　三ツ木萌 §1

[1] 明星大学情報学部

## 1　はじめに

自動運転をはじめとする様々なシステムで画像認識技術が活用されており，近年では特に Convolutional Neural Network（CNN）によるものが一般的に用いられている．実システムにおいては，リアルタイム，高精度かつ低消費電力な画像認識器を実現する必要があり，Field-Programmable Gate Array（FPGA）を活用することでこれらの要求を満たすことが可能となる．

本稿では，FPGA 上に実装された CNN による画像認識器を対象とし，動的に学習結果を反映し精度を継続的に向上させる手法を実現したシステムの性能評価を行い，その有効性を示す．

## 2　関連研究

ディープラーニングなどのニューラルネットワークを用いた推論技術はクラウド基盤上に構築されたフレームワークを活用することが多いが，クラウド基板上で推論を行うためには，処理遅延に加え，通信量やデータ容量の増大による通信の遅延等も考慮する必要がある．そのため，リアルタイムな推論処理が難しくなるとともに，システム信頼性・頑健性の低下，電力消費増大などの問題が生じる．そこで，実際の学習・推論対象となるセンサデータを取得する IoT デバイス上で推論処理等をリアルタイムに実行するエッジコンピューティングが注目されている．

鈴木らの研究では，処理の遅延や通信量の増大が問題となっているストリーム型自動運転システムに対して，LDM(Local Dynamic Map) を用いて処理の遅延時間を短縮している [3]．

本稿は FPGA に実装したニューラルネットワークの重みパラメータ等を動的に更新し，推論精度を継続的に向上させる仕組みを実現することを目的としており，より広い適用範囲を目指したものである．

## 3　CNN パラメータ動的更新手法

エッジデバイスでニューラルネットワークによる推論をリアルタイムに行うためには，ネットワークの深さや規模を削減してネットワークの構造を単純化する，パラメータの精度を削減して演算を高速化する，といった手法が考えられる．しかしながら，単純化・小型化したニューラルネットワークは複雑・大規模なニューラルネットワークと比較して，学習・推論の精度が低下する傾向にある．エッジデバイスの演算性能不足を補い，認識精度を継続的に向上させるために，筆者らは従来より，サーバで学習を行う学習オフロードシステムと FPGA を備えたエッジデバイスを連携させ，常に最新の学習データを用いた高精度・高速な推論を可能とするシステムを提案している [2, 1]．このシステムでは，FPGA を備えたエッジデバイスとサーバは同じニューラルネットワーク構造を使用して推論・学習を行い，サーバで行われた学習の結果を取得して FPGA に反映することで，常に最新のパラメータを用いた推論を可能とする．

FPGA ボードはサーバとネットワークを介して最新の学習結果（パラメータ）を取得する．取得されたパラメータは DRAM を介して，ニューラルネットワークを実行している FPGA 内 Block RAM に送信される．これにより FPGA は回路情報を更新することなく，動的に推論に用いる学習結果を更新することが可能になる．

推論パラメータ動的更新の手順を図 1 に示す．FPGA のプログラマブルロジック（PL）は，起動時にプロセッシングシステム（PS）からの推論パラメータの更新をフラグを使用してチェックする．更新が無い場合は Block-RAM に保存されている推論パラメータを用いて推論

---

* Performance Evaluation of a Dynamic Update Method for CNN Weight Parameters on an FPGA

† Musashi AOTO, Meisei University

‡ Yasutaka WADA, Meisei University

§ Moe MITSUGI, Meisei University

を開始する．推論パラメータに更新があった場合は，DRAM から更新パラメータをロードし，BlockRAM にロードする．以後は，更新がない場合と同様 BlockRAM の推論パラメータをロードして推論を行う．

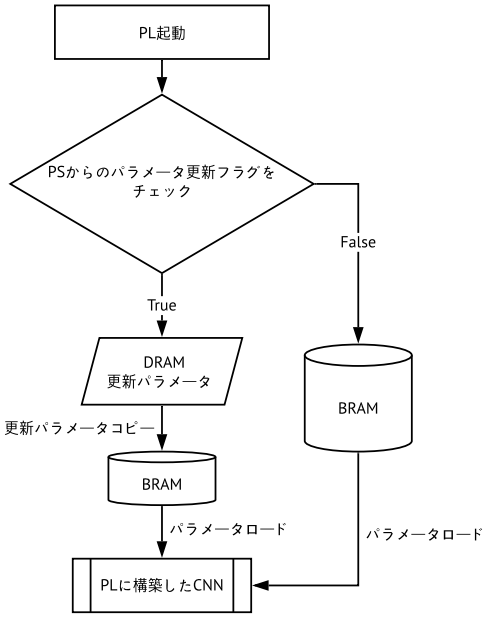


図 1　FPGA 上の CNN パラメータ更新フロー

## 4　CNN パラメータ動的更新手法の性能評価

本評価では，画像認識に用いられる InceptionResNetV2 を Keras にて実装したもの [1] を用い，認識精度の推移を学習オフロードの有無によって比較した．なお，学習のパラメータを表 1 に示す．

本評価の結果を図 2 に示す．図 2 では，横軸が学習結果の反映回数を，縦軸が学習に用いた画像データの数および認識精度を表している．図 2 より，学習処理のオフロードにより学習に用いるデータを追加・蓄積することができるとともに，より多数のデータを用いた学習を通じて，学習結果を随時反映するパラメータ動的更新手法により，継続的に認識精度が向上していることがわかる．

## 5　まとめ

本稿では，FPGA に実装したニューラルネットワークの重みパラメータ等を動的に更新し，推論精度を継続的に向上させる仕組み [1] を実現し，その性能評価を行なった．ネットワークを介してエッジデバイス上の DRAM に転送された学習済みデータを，FPGA 内部の

表 1　InceptionResNetV2 学習パラメータ

| | |
|---|---|
| エポック数 | 1,000 |
| ミニバッチサイズ | 64 |
| 学習率 | 0.1% |
| 画像サイズ | 360 * 240 [pixels] |
| 学習データ数 | 12,700 |
| 検証データ数 | 1,000 |

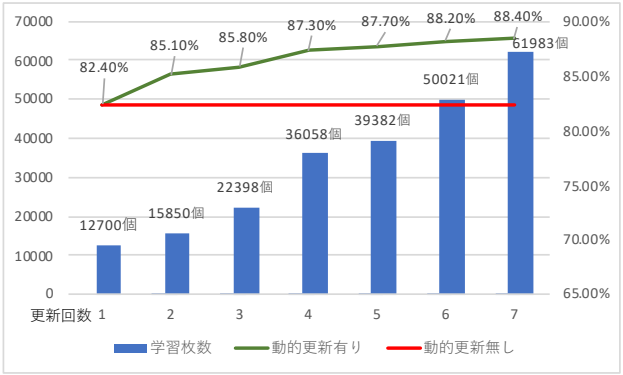


図 2　CNN 重みパラメータを動的に更新する手法を用いたシステムの性能評価

BRAM に読み込むことでニューラルネットワークの重みパラメータを動的に更新する．性能評価の結果，高性能サーバに学習処理をオフロードし，その結果を活用することで，エッジデバイスによる推論の精度をより高速に向上させられることが確認された．

今後の課題として，重みパラメータだけではなく，ニューラルネットワークモデルも含め動的に更新し，複数のニューラルネットワークをシームレスに切り替えて動作するシステムの実現が挙げられる．

**謝辞**　本研究の一部は JSPS 科研費 17K12665 および 18K19786 の助成を受けたものです．

## 参考文献

[1] Musashi Aoto, et al. Towards the improvement of training efficiency and image recognition accuracy for an fpga controlled mini-car by offloading neural network training. In *Proc. of FPT*, pp. 437–440, 2019.

[2] 青戸武蔵ほか．単機能なニューラルネットワークを複数用いた高速・高精度な画像認識の FPGA による実現．信学技報, Vol. 119, No. 208, pp. 57–62, 2019.

[3] 鈴木有也ほか．クラウド型自動運転を指向したストリーム処理型 LDM の低遅延処理手法．組込みシステムシンポジウム 2015 論文集, pp. 84–92, 2015.