FPGA を用いたリアルタイム AI カメラシステムの設計と評価

Design and evaluation of a real-time AI camera system using FPGA

1235119　亀阪 亮紀 （Soft Intelligent System on Chip 研究室）
（指導教員　星野 孝総 准教授）

1. はじめに

カメラ画像を人工知能（以下，AI）システムで多様な画像分類タスクを行う場合，一般的に高性能な計算機を必要とし，GPU を用いたアプローチが一般的である．GPU を用いた場合，高速なメモリ確保・転送の仕組み，高電力源の確保を必要とし，実用的な IoT 向け機器への実装は，リアルタイム性や省電力化への課題がある．画像認識分野で注目されている畳み込みニューラルネットワーク（以下，CNN）の演算は積和演算の集合であり，繰り返し処理ではなく流れ演算である．そこで，本研究では，GPU を用いずに，FPGA（Field Programmable Gate Array）を用いてカメラからの転送速度に合わせてパイプライン処理を行い，リアルタイム同期処理の実現を目指す．これにより，カメラデバイスと AI 処理が一体となったハードウェアデバイスとして用いることができ，メモリに格納する前に AI 処理を部分的に終了させ，CPU 側の処理を極端に軽減することが期待できる．本研究では，AI 搭載型害獣捕獲システム[1]をターゲットアプリケーションとした．

2. リアルタイム AI カメラシステム

本研究では，カメラが N 番目の画像 1 フレームの転送が終了する前に，N-1 番目の処理を終えていることをリアルタイム処理と定義した．このようにすることで，カメラのスルーレートを守りながらフレーム落ち無しに処理をしている状態になる．本研究では，図. 1 に示すようにカメラの取り込み終了時から，処理が完了するまでの時間を計測した．

図. 2 に設計したカメラシステムの概要を示す．カメラから転送されてくるデータは FIFO に取り込まれる．FPGA 側(動作周波数：200MHz)では FIFO から取り出させたデータを画像処理 IP へ転送し，処理を行い，DDR3 SDRAM に DMA 転送する．DMA 転送された画素データを元に生成された画像は Ethernet を経由して UDP 通信で転送を行い，クライアント側で受信し，表示させることで確認を行った．

3. 実験内容

3.1 カメラ信号を用いた畳み込み層のレイテンシー計測実験

FPGA 上で 2 層の畳み込み層を設計し，カメラ信号を用いたレイテンシー計測を行った．本実験では，3x3 の畳み込みと 2x2MAX プーリング行う畳み込み層を 2 層用意し，SXVGA 規格の画像に対して，レイテンシーの計測を行った．

3.2 全結合層に向けた多層ニューラルネットワークの設計

CNN の全結合層に向けた FPGA ベース多層ニューラルネットワーク（以下，NN）の設計を行った．設計方法として，本研究のターゲットアプリケーションに用いられている画像データセット[1]を用いて整数型 2 クラス分類器を作成した．NN のモデルは，入力 784，隠れ層 1 のニューロンの個数が 60，隠れ層 2 が 20，出力 2 の構造である．

3.3 背景差分法を用いた対象物体のラベリング

CNN の前処理として，背景差分法による物体検出[1]を FPGA 上に実装した．背景差分法による二値化画像を生成し，その二値化画像に対して，ランレングス圧縮を行い，その後，結合比較処理を行い，ラベリングを行った．

4. 実験結果と考察

第 3.1 章の実験結果を表. 1 に示す．実験結果より，1 層目と 2 層目合わせて合計 10 ラインの遅れとなり，次フレームの約 10 ラインが取り込み終了時に 2 層の畳み込み層での処理が行われていることになる．1 フレーム 960 ラインであるため，約 1%程度の遅れとなっている．つまり，次フレームの取り込みが完了するまでに処理が完了していることになり，アルタイム画像処理が実行できている．

第 3.2 章の実験結果を表. 2 に示す．ニューロ演算を並列処理なしで行った場合，積和演算器をループ型逐次処理で使用しているため，特徴量が多い 1 層目が推論時間の殆どを占めている．1 層目の積和演算器を 2 並列化した場合，理論値通り，ループ型逐次処理に比べ，半分の推論時間となった．

第 3.3 章の実験結果を表. 3 に示す．背景差分法とランレングス圧縮はラインバッファへの格納する必要が無く，ほぼレイテンシーをゼロで行うことができている．対して，ラベリング処理は On Chip RAM をループ処理で比較しているため，他の処理に比べ，レイテンシーが大きくなっている．

5. おわりに

本研究では FPGA を用いてリアルタイム AI 処理の実現を目的とし IoT 向け AI カメラシステムの開発を行った．FPGA によるレジスタ転送レベルでリアルタイム AI 処理ができている事を確認でき，ハードウェア実装することの有効性を示すことができた．

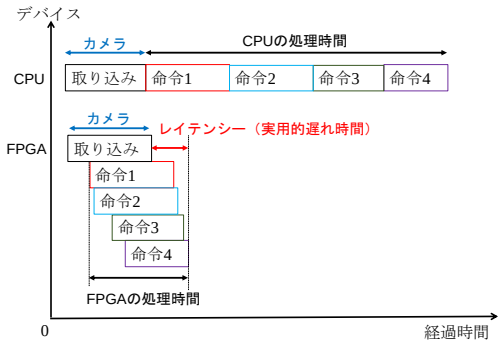

図. 1　本研究におけるレイテインシーの定義

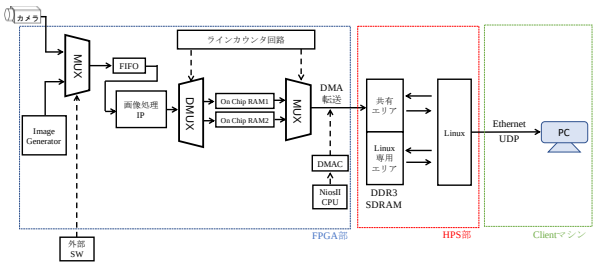

図. 2　ハードウェアベース処理システムのアーキテクチャ

表. 1　畳み込み層の各層におけるレイテンシーの計測結果

Device	Layer1[us]	Layer2[us]
FPGA	44.7	22.5

表. 2　4 層 NN による画像 1 枚あたりの推論時間の計測結果

Device	逐次処理[us]	1 層目 2 並列化[us]
FPGA	241.73	123.98

表. 3　背景差分から結合比較のレイテンシーの計測結果

Device	背景差分[us]	ランレングス圧縮[us]	結合比較[us]
FPGA	0.01	0.18	9.20

参考文献

[1] 亀阪亮紀．畳み込みニューラルネットワークを用いた害獣捕獲システムの試作と検討．高知工科大学 卒業研究報告書, 2019.