平成30年度 情報工学コース卒業研究報告要旨

高田研究室 研究室	氏　　名	稲 石 日 奈 子
卒業研究題目	機械学習の推論ハードウェアの C言語設計における性能見積もり手法	

近年，深層学習は物体認識や領域分割，姿勢推定，深さ推定など様々な分野で利用されている．深層学習には学習と推論のフェーズがあり，それぞれの特徴に応じて異なるデバイスが使用されている．学習では大量のデータを使用し，何度も演算を繰り返す必要がある．そのため，電力消費は大きいが，学習において高い性能を持つ GPU が使用されることが多い．一方推論では用途に応じて消費電力やコスト，応答時間など様々な要件があり，それに応じて GPU や CPU，FPGA，専用ハードウェアが使用されている．GPU は機器への組込みという観点では，電力効率や容積などの制約が多い．また， FPGA は特定の演算に対して CPU より高速に演算を行うことができ，かつ専用ハードウェアと比べて柔軟に設計できる．そこで本研究では，研究室におけるシステムレベル設計ツールを用いた C 言語設計による，効率の良い FPGA 向けの推論器のための C ソースコード記述の検討を行った先行研究を受けて， FPGA で行う推論に着目している．先行研究では，6 レイヤの CNN（Convolutional Neural Network）に対して 8 つの高速化手法を検討しており，それぞれの高速化手法は，推論器全体に適用されるものとレイヤごとに適用されるものがある．例えば 6 レイヤの CNN の各レイヤに，ある 1 つの手法を適用するしないを検討する場合，2^6 通りの選択肢があることになり，この組み合わせの数は NN の規模によって膨大な数になる．そのため，どの高速化手法を用いるとより効率的に推論を行うことができるか評価するために，適用する手法の組み合わせを変更して何度も高位合成や論理合成を行う必要がある．しかし，高位合成や論理合成には時間がかかるため，ハードウェア設計にかかるコストが大きくなる．

そこで本研究では， 高位合成および論理合成をすることなく，性能を見積もる手法を検討した．そのために各レイヤの実行サイクル数をあらかじめ取得し，それを参照値として，パラメータを用いて見積もり式を検討する．パラメータは入出力画像サイズ，入出力画像枚数，カーネルサイズ，ストライドとする．実行サイクル数の取得方法は次の 2 つの方法を検討した．1 つ目は高位合成ツールが生成した合成レポートから取得する方法，2 つ目はある特定のパラメータを持つレイヤのシミュレーション結果から取得する方法である．本研究では評価対象として，先行研究で検討されている高速化手法のうち，各レイヤごとにそれぞれプロセス分割し，レイヤレベルでパイプライン実行する方法，次のレイヤに送る特徴量画像データを FIFO で送信する方法，重みデータを複数個まとめて 1 つのデータとして送受信を行う方法が適用されているソースコードを用いた．

高位合成ツールが生成する合成レポートを用いて見積もりを行う方法は，条件分岐に関する情報が不足しており，また最適化の情報がレポートに反映しきれていなかったため，困難であるとわかった．しかし，各レイヤの個別のシミュレーション結果から実行サイクルを取得し，見積もり式を検討した方法では，レイヤ全体の入力画像サイズを変更した場合の計測値と見積もり値の誤差が 10%程度で見積もることができ，満足できる結果を得ることができた．特に誤差の大きいパラメータ条件で詳細を確認したところ，個別レイヤでの実行サイクル数取得時に入力画像サイズに依存した原因があることが判明した．今後はこの原因の解明と見積もり式の再検討を行い見積もりの精度向上を目指す．また，本論文で評価対象とした高速化手法はごく一部であり，ループ内の処理をパイプライン化するフォールディングやループ展開など他の高速化技術が適用されたプログラムに対しても見積もりを行う予定である．