

福岡大学  
博士学位論文要約

RESEARCH ON LOW POWER CNN IMPLEMENTATION WITH  
APPROXIMATE ARITHMETIC FOR EDGE INFERENCE

近似演算器を用いたエッジ推測向けの低消費電力  
畳み込みニューラルネットワークの実装に関する研究

(平成31年3月)

楊 同鑫

工学研究科 情報・制御システム工学専攻

最近、人工知能が自動運転、癌検査、工場の自律ロボットなど、様々な分野に使われている。その中で、最も利用されているのが Convolutional Neural Network (CNN) である。CNN は高い認識率を持つため、特に膨大な計算の必要な画像認識などのアプリケーションによく使われている。CNN の学習は大量のデータに対して膨大な計算処理を行うので、クラウドやサーバーに学習モデルを作成して学習し、エッジ側の組み込みシステムでは推論のみを実施するのが一般的である。エッジ側で CNN の推論をする理由には、セキュリティ、プライバシー、そしてレイテンシなどの考慮をしなければならないことも含まれる。しかし、エッジ側の組み込みシステムは面積や電力供給が限られることが問題で、CNN を様々な分野に利用できるようにするためには、これらの課題を早急に解決する必要である。

本博士論文では、近似演算を用いたエッジ推論向けの低消費電力な CNN を実装することを目的として行った研究について記した。

第 1 章では、本研究の動機を説明した。CNN では大量な畳み込み演算が必要であり、その電力を削減することがとても重要である。CNN は多数の畳み込み演算回路を利用して同時に演算するため、畳み込み演算回路の面積を小さくできれば小面積の CNN エッジデバイスが得られる。一方、CNN は推論の精度を高めるため、従来の畳み込み演算では 32 ビットの浮動小数点の乗算と加算を利用している。しかし、32 ビット浮動小数点演算は電力も回路面積も大きく、エッジ側の推論には非効率である。少ビット、かつ、固定小数点の演算は電力と面積の削減に効果がある。また、近似計算は演算結果に多少誤りを持つことで、電力と面積が削減できる。そこで、近似計算の考えを少ビットの固定小数点の演算に導入すれば、省電力と小面積を同時に実現できる。多少の誤差があったとしても、その結果が CNN のアプリケーションで許容できれば問題ない。更に、CNN は異なるレイヤーにおける演算精度の要求が異なるので、この特性を利用して消費電力の削減率を最大化できるように、動的に精度可変な近似演算器が有効であると考えられる。

第 2 章では、本研究の背景を明らかにするために、代表的な CNN アーキテクチャと CNN 研究のための環境を記述した。その後、現状のエッジ推論と近似計算の関連研究について記した。

第 3 章では、本博士論文のこれからの各章の提案の評価方法を記述した。消費電力、面積、遅

延時間, そして精度の評価方法を説明した.

第4章では, 乗算器の部分積ツリーを効率よく圧縮する Approximate Tree Compressor (ATC) を提案した. この ATC を用いる 8 ビット乗算器は, 従来のツリー型 (Wallace Tree) 乗算器と比べて, 消費電力で 60%, 面積で 50.1%小さくなった. 近似乗算器単体での精度を評価し, その低下が小さいことを確認した. 実用面の評価をする目的で, CNN でよく使われるものと同じ  $5 \times 5$  のサイズのフィルタを用いた画像の鮮鋭化処理で, 提案する近似乗算器の精度を評価した. その結果からこの画像処理アプリケーションで問題なく利用できることを確認した.

第5章では, 近似計算と正確な計算を動的に切り替えることができる桁上げをマスク可能な加算回路を提案した. この加算回路を利用して 16 ビットの Carry-Maskable Adder (CMA) を実装し, 従来の ripple carry adder (RCA) と比較した. CMA に 4 種類の精度を設定して電力を評価し, 最大 54.1%の電力を削減できることを確認した. 更に, 画像の鮮鋭化処理アプリケーションにより, 精度可変性の有効性を確認した.

第6章では, 上記の ATC と CMA を使って, 精度可変の近似乗算器を構成した. この評価結果を分析し, 問題ない範囲で更に精度を落とすことで, 電力を一層削減できることを見出した. その改良に近似符号処理の回路を追加して, 符号付き数値に対応する動的精度可変な近似乗算器 Efficient Accuracy-Controllable Multiplier (EACM) を提案した.

第7章では, 代表的な CNN である LeNet-5 を取り上げ, EACM の消費電力と精度を評価した. 符号付き数値を処理できる Baugh-Wooley アルゴリズムを利用した Wallace Tree 乗算器を比較対象とした. EACM は面積を 61.2%削減できた. 上記二つの乗算器を使って, それぞれ LeNet-5 を実装した. 手書き文字認識用の MNIST データセットを使って比較した実験結果から, 演算精度の設定により EACM は 30.5%~42.6%の消費電力を削減できることが確認された. 95.8%~97.2%の高い文字認識率を保持していることも確認できた. Wallace Tree 乗算器を利用する場合の 97.4%の認識率と比べて, 高い演算精度を設定した際の EACM での認識率の差は 0.2%と非常に小さく, 低い演算精度を設定した際でもその差は僅か 1.6%である.

第8章では, 本論文の結論と将来の展望について記述した.