氏　　名	よう　どうきん 楊　同鑫
学位の種類	博士（工学）
報告番号	甲第1754号
学位授与の日付	平成31年3月14日
学位授与の要件	学位規則第4条第1項該当（課程博士）
学位論文題目	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）
論文審査委員	（主　査）　福岡大学　教授　佐藤　寿倫 （副　査）　福岡大学　教授　中西　恒夫 九州大学　教授　井上　弘士

内　容　の　要　旨

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 × 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い出した．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．

審査の結果の要旨

審査経過

1. 平成 30 年 11 月 21 日の博士論文事前審査会で，審査者は申請資格条件に適合すると判定された.
2. 平成 30 年 10 月 29 日の類似度判定で，申請論文は学位論文として問題無いことが確認された.
3. 平成 30 年 12 月 3 日の第 1 回審査会で審査論文に対して質疑と指示があり，公聴会の開催が了承された.
4. 平成 31 年 1 月 15 日の公聴会と最終審査会で，申請論文は博士（工学）の学位論文に値すると認められ，申請者は合格と判定された

審査委員の結論

スマートフォンなどの携帯機器や IoT（Internet of Things，モノのインターネット）デバイスが次々と現れており，中でも AI（Artificial Intelligence，人工知能）デバイスへの期待は大きい．これらの普及には，機器やデバイスが消費する電力の削減が必須である．一方で，半導体技術の進展に伴い，電源電圧を低下させることによる従来ながらの消費電力削減が困難になってきた．以上の背景の下，LSI デバイスの消費電力を削減する新たな技術が求められていた.

この課題に対して，申請者のヨウ・ドウキン氏は近似計算（Approximate Computing）による解決を提案している．現在は性能・面積・消費電力の間のトレードオフを考慮して LSI デバイスの設計が行われているが，計算精度という軸を加えることでトレードオフの探索空間を拡大し，消費電力を削減することを検討している．シミュレーションなどの科学技術計算とは異なり，画像処理やディープ・ラーニング応用では必ずしも高い演算精度は要求されない．また，センサにより獲得されたデータにはそもそもノイズが混入されており，高精度での演算に意味が無い場合もある．AI デバイスが対象とするアプリケーションにはこのような性質があり，ヨウ氏はこの点に着目して，エッジデバイスにおける CNN（Convolutional Neural Network，畳み込みニューラルネット）での推論処理に特化した消費電力削減技術を考案している.

具体的には，CNN で多用される積和演算に着目して，消費電力と回路面積を大きく削減できる近似乗算器と，精度を演算処理時に変更可能な近似加算器とを考案し，積和演算器を実現している．代表的な CNN である LeNet をディジタル回路として設計し，手書き文字認識アプリケーションの処理時における消費電力を評価している．文字認識精度を 97.4%から僅か 1.5%低下させるだけで，消費電力と回路面積をそれぞれ 42.6%と 62.8%削減することを可能にしている．IoT や AI 応用分野において実用性と波及性の高い研究成果であり，情報・制御システム工学への寄与が大きい．よって，本論文は博士（工学）の学位論文に値すると認める.

本論文に対し審査委員から質問や指摘があったが，ヨウ氏から的確な回答がなされた．公聴会での聴講者からの質問でも，ヨウ氏の説明により質問者の理解が得られた．以上の結果，ヨウ氏は試験に合格したものと認める．