別紙様式８

研　究　主　論　文　抄　録

論文題目
再構成性をもつ DNN アクセラレータに関する研究

熊本大学大学院自然科学教育部　工学専攻　先端情報通信工学教育プログラム
（　主任指導　飯田　全広　教授）

論文提出者　中原　康宏

主論文要旨

近年，DNN（Deep Neural Network）が画像認識や音声認識などの分野で非常に高い成果を上げており，多くのアプリケーションに導入されている．しかし，DNN の計算コストは高く，処理を行うデバイスの電力消費が DNN を処理する際の課題となっている．
そのため，DNN 処理を目的とした様々なデバイスが提案されている．これらのデバイスは主に 2 つの手法で消費電力の削減を行っている．1 つ目は DNN 処理に合わせて処理要素やデータ転送経路の最適化する手法である．2 つ目はパラメータや入力データなどのビット幅を削減することで，1 つのデータを処理するのに必要なコストを削減する手法である．
しかし，これらの手法には課題が存在する．前者の手法は特定の演算に構造を最適化してしまうため，それ以外の演算を処理する場合は処理効率が低下するという課題がある．DNN はそのモデルごとに様々な構造やハイパーパラメータをもつため，これらすべてに 1 つの構造で対応することは難しい．後者の手法は演算に必要なコストの削減と引き換えに推論精度が低下するという課題がある．精度低下を防ぐために再学習と呼ばれる手法も存在するが，一定上の精度低下は阻止することが出来ず，再学習に必要な設備，時間的コストは大きい．また，これらのデバイスの多くはスケーラビリティに乏しいという課題も存在する．DNN モデルには大小さまざまなモデルが存在する．大規模なモデルを処理する場合，演算リソースが限られるこれらのデバイスは多くの処理時間を必要とするため要求される性能を満たせない．データのビット幅削減によりこの問題を緩和したデバイスも存在するが，この手法にも限界があり，大規模なモデルに対応することは難しい．
本研究では，以上の課題を踏まえて様々な DNN を低消費電力かつ高速に処理可能な DNN アクセラレータについての検討を行う．まず，多種多様な DNN に対応するために，再構成性をもつ DNN アクセラレータについて検討を行う．再構成性とは処理内容に合わせて処理要素が行う演算やデータ経路などを任意に変更可能な特性であり，この再構成性により 1 つの回路で複数の DNN を効率的に処理可能になることが期待できる．次に，データのビット幅と推論精度のトレードオフを改善するために，新たな演算器についての検討を行う．

本演算器は近年提案された数値表現 Posit に対応した演算器である．DNN 処理において Positは低ビット幅と高推論精度を両立可能な数値表現であるが演算器の回路面積が大きいという課題をもつ．この演算器を小面積に改良することで低ビット幅，高推論精度かつ省面積な DNN アクセラレータの実現が期待できる. 最後に複数の DNN アクセラレータを用いた高速化について検討を行う．これにより，モデルの規模に対するスケーラビリティを確保することができ，先述した再構成性と併用することで様々なモデルや要件に対応できることが期待できる．

本論文は 6 章から成り，第 1 章では本研究の背景と目的について述べる．

第 2 章では，DNN アクセラレータについて説明を行い，関連用語の定義を行う．その後，昨今の DNN アクセラレータについて説明し，本研究の趣旨を以下のように位置付ける．

・回路再構成による様々なハイパーパラメータをもつ DNN モデルへの対応
・新たな積和演算器によるデータの低ビット幅，高推論精度，小面積の両立
・複数のデバイスを用いた DNN のアクセラレーション

第 3 章では再構成性をもつ DNN アクセラレータ ReNA のアーキテクチャの提案を行った．DNN の一種である CNN（Convolutional Neural Network）は大きく畳み込み層と全結合層の 2 種類の層からなる．ReNA はデータの転送経路の再構成性により，この 2 種類の層を単一の回路で効率的に処理することができる．本研究では，代表的な DNN モデルである AlexNet と VGG を実行した場合の ReNA の性能について評価を行った．アクセラレータの実装には TSMC 22nm のスタンダードセルプロセスを用いた．評価の結果，高い電力効率を達成した．

第 4 章では，数値表現 Posit に対応した新たな積和演算器の提案を行った．Posit は DNN 処理においてデータの低ビット幅と高推論精度を両立した数値表現である．しかし，Posit の積和演算器は構造上値を累積する部分が大きくならざるを得ず，回路面積が大きい．消費電力は回路面積に比例する傾向にあるため, 消費電力がネックとなる場合 Posit の導入は難しい．そこで，新たな累積の手法を提案し，累積部分の面積削減を行った．回路の実装には，TSMC 22nm のスタンダードセルプロセスと FPGA（Intel Stratix 10）を用いた．8 bit の Posit 演算器について評価を行った結果，同等の推論精度保ちながら ASIC で 43%の面積削減，FPGA で 63%のリソース使用量削減を達成した．

第 5 章では，複数台の FPGA を用いた DNN 処理のアクセラレーションを提案した．従来の DNN アクセラレータは単一のデバイスで処理を行うものが多く, 処理速度はデバイスの演算資源に律速される．そこで，本研究では複数台のデバイスを用いたアクセラレーションを提案した．再構成性をもつ DNN アクセラレータ ReNA を FPGA 向けに改良し，複数台の処理に対応させることで再構成性とスケーラビリティを両立する．本アクセラレータの実装には FPGA（Intel Stratix 10）を用いた．VGG16 を対象に評価を行い，高い処理速度を達成した．

最後に，第 6 章で本研究の成果についてまとめ，今後の課題を述べる．