様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２１ 日現在

機関番号：３２６７８
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１８８５

研究課題名（和文）メモリサイズ削減を目指した融合型ニューラルネットワークアクセラレータの開発

研究課題名（英文）Fused-layer neural network accelerators with reduced on-chip memories

研究代表者
瀬戸　謙修（Kenshu, Seto）

東京都市大学・理工学部・講師

研究者番号：１０４２０２４１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：ニューラルネットワークの複数のレイヤを並列実行するハードウェアである融合型ニューラルネットワークアクセラレータでは、中間結果をオンチップメモリに保存することでオフチップメモリアクセスを削減することができ、その結果、消費エネルギー削減効果を期待できる。本研究課題では、融合型ニューラルネットワークアクセラレータのアーキテクチャ最適化に有効な、ループ最適化やメモリアクセス最適化技術の開発に成功した。メモリアクセス最適化では、自動ラインバッファ化で有効なスカラリプレイス技術について、複数の書き込みアクセスが扱えなかった課題を解決した。

研究成果の学術的意義や社会的意義
AIの基盤となるニューラルネットワークを効率良く実行するハードウェアが求められており、融合型ニューラルネットワークアクセラレータに着目した。アクセラレータを短期間に設計するには、プログラムからハードウェアを自動生成できる高位合成の活用が効果的だが、高効率なハードウェア生成には、人手によるコード最適化が必要になり、設計期間の長期化が問題となっていた。重要なコード最適化の一つとして、スカラリプレイスと呼ばれるメモリ最適化があり、これを融合型ニューラルネットワークに適用できるよう、世界初の拡張に成功した。研究成果の活用により、迅速な高効率ニューラルネットワークアクセラレータの設計に寄与できる。

研究成果の概要（英文）：Fused-layer neural network accelerators, which execute multiple layers in parallel, reduce off-chip memory accesses by storing intermediate results in on-chip memories. As a result, we expect reduced energy consumption for fused-layer neural network accelerators. In this research, we successfully developed loop and memory access optimizations which are effective for developing fused-layer neural network accelerators. For the memory access optimization, we extended the scalar replacement, which is useful for automatic line buffer generation, so that it can handle multiple write accesses.

研究分野：高位合成、並列化コンパイラ、ハードウェア・ソフトウェア協調設計、VLSI設計技術

キーワード：ニューラルネットワーク　アクセラレータ　オンチップメモリ削減　高位合成

１．研究開始当初の背景

ニューラルネットワーク(以下、NN と略記)は、高精度な画像認識などの鍵となる技術である。クラウドへの通信量の急激な増加を防ぐためなど、モバイル端末やセンサノードなどのエッジデバイスでの NN 推論が求められている。リソース制約の厳しいエッジデバイスで NN 推論を効率的に実行するため、様々な NN アクセラレータが提案されている。4K など超高解像画像に対する推論や、ピクセル単位の推論を行うセグメンテーションなどの高負荷な NN アプリケーションをエッジデバイスで実行するには、NN アクセラレータの一層の高速化とともに、低エネルギー化の両立が必要である。

NN の各レイヤを順に実行する NN アクセラレータにおいて、レイヤ間で受け渡される画像データ(特徴マップ)をオフチップメモリに格納する場合、大量のオフチップメモリアクセスが問題となる。この問題を解決するため、融合型 NN アクセラレータが提案された。融合型 NN アクセラレータは、複数のレイヤをチップ上でパイプライン実行し、レイヤ間の画像データをオンチップメモリ経由で通信することで、オフチップメモリアクセスを削減する。オンチップメモリサイズを更に削減するため、レイヤ間の画像データの受け渡しにラインバッファを使用する方法が提案された。例えば、フィルタカーネルサイズが 3×3 の畳み込み(Conv)レイヤの場合、4 ライン分の入力バッファ、および、2 ライン分の出力バッファを使用する。しかしながら出力バッファは、入力バッファと同一のデータを保持しているため、冗長である。そこで、融合型 NN アクセラレータについて、性能を維持したまま、一層のバッファメモリ削減手法を検討する。

２．研究の目的

リソース制約やエネルギー制約の厳しいエッジデバイス向けに、バッファメモリサイズを最小限に小さくした、省メモリな NN アクセラレータのアーキテクチャを提案するとともに、そのアーキテクチャに基づいた NN アクセラレータの設計手法を明らかにすることで、迅速な NN アプリケーションのエッジデバイス上での最適実装を支援することを目的とする。

３．研究の方法

高効率な NN アクセラレータを設計するため、畳み込みレイヤ、プーリングレイヤなど、NN を構成するレイヤのリファレンス C コードをベースとし、ループ最適化、メモリアクセス最適化などの高位合成向けソースコード最適化の適用を検討した。レイヤ間の接続には、ストリームインターフェースを活用する。複数のレイヤをパイプライン実行する際、正しい動作のために、隣接する N 番目、N+1 番目のレイヤの間で、N+1 番目のレイヤによる入力データのアクセス順と、N 番目のレイヤが出力するデータのアクセス順を一致させた。

ループ最適化では、ループ融合、ループ展開とループ平坦化の検討を行う。ループ融合では、ループ交換やループシフトも適用しながら、複数のループを一つにまとめることで、並列性やメモリアクセス局所性の向上を行うことができる。ループ融合やループ展開を適切に行うことで、スカラリプレイスと呼ばれるメモリアクセス最適化が適用可能になり、メモリアクセス競合の解消とともに、オンチップメモリのサイズ削減につながる。また、NN のレイヤは多重ループとして記述できるが、多重ループのまま高位合成を行う場合、ループをまたぐ際に実行サイクル数のオーバーヘッドが発生する。このオーバーヘッドを削減するのに有効なループ平坦化について検討を行う。メモリアクセス最適化では、特徴マップを格納する配列に対してスカラリプレイス、重みを格納する配列に対してメモリ分割について検討を行った。ループ融合やスカラリプレイスでは、多面体モデルと呼ばれるループプログラムの表現形式を用い、多面体モデルの操作が可能なライブラリである Integer Set Library(ISL)を使用した。エッジデバイス向けの NN 実装では、プルーニングによる NN 自体のコンパクト化も重要であると考え、検討を行った。高位合成向けソースコード最適化の適用後には、高位合成ツールを実際に使用して、生成された回路の実行サイクル数や回路面積の評価を行った。

４．研究成果

高位合成を使用して、リファレンス C コードから高効率な NN アクセラレータを設計する際、ループ融合やループ平坦化などのループ最適化が有効である。多面体モデルを用いたループ融合では、実行文ごとに多次元スケジュールを求めた後に、コード生成を行う。コード生成のアルゴリズムによって、分割されたループが出力されてしまう場合があるため、ループの分割を防ぐ工夫を提案し、国内会議で発表した。また、広く用いられているコンパイラフレームワークである LLVM をベースとしたループ融合ツールを開発することで、従来技術と比べて広い範囲の C コードを受け付けられるように機能拡張した。さらに、従来手法でループ融合する際、多次元スケ

ジュールに対する後処理を行うが、融合後のループ回数が不必要に増加する課題があったため、ループ回数を削減する方法を提案した。ループ平坦化では、多重ループを一重ループにする際、一重ループのループ変数から、元のループ変数を復元する必要があるが、その復元処理を低オーバーヘッドで実行する方法を提案し、国内会議で発表した。多面体モデルを用いたループ最適化の最適化空間を探索するには、モンテカルロツリー探索が有効であると考え、スケジューリング手法について研究し、国内会議で発表した。さらに、NN のプルーニングにおいて、各レイヤの重み削減率を決定する方法を考案し、国内会議で発表した。

ループ融合によるメモリアクセス局所性向上を、高位合成を用いたハードウェア設計に活かすには、スカラリプレイスと呼ばれるメモリアクセス最適化が必要となる。スカラリプレイスにより、メモリアクセス競合を削減して処理時間を短縮するとともに、オンチップメモリを削減できる。しかし、NN のリファレンス C コードにおける出力特徴マップに相当する配列など、例えば初期化を伴う配列の場合、ループ融合の結果、ループ本体中に配列への書き込みアクセスが複数回発生し、従来手法では、スカラリプレイスが適用できなかった。そこで、配列への書き込みアクセスが複数回ある場合でもスカラリプレイスを行える方法を提案した。その際、同時に削除可能な配列アクセスを網羅的に列挙し、開始間隔の制約条件を満たしつつ、発生するシフトレジスタ数を最小化する方法も提案、ツール実装し、国際会議で発表した。ループ融合後のベンチマークコードに対して、提案するスカラリプレイスを適用し、高位合成により評価した結果、最適化前の設計と比べて、9%のゲート数の増大に対し、性能を平均 2.1 倍スピードアップすることに成功した。本研究成果は、エッジデバイス向けの省メモリな NN アクセラレータ開発において世界初の技術であり、必須の設計技術の一つになると期待される。さらに、NN アクセラレータの高位合成を用いた設計にて効果的なループ最適化やメモリアクセス最適化をまとめ、招待論文（査読付き英文論文誌）にて発表した。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Kenshu Seto	４．巻 16
２．論文標題 A Survey on System-Level Design of Neural Network Accelerators	５．発行年 2021年
３．雑誌名 Journal of Integrated Circuits and Systems	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.29292/jics.v16i2.505	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Yuta Hiyama, Takayuki Todokoro, Kenshu Seto, Masato Tatsuoka, Yoshihito Nishida, Mineo Kaneko
２．発表標題 High-level synthesis code optimization with loop fusion based on LLVM/Polly
３．学会等名 The 22nd Workshop on Synthesis And System Integration of Mixed Information Technologies (SASIMI) 2019 (国際学会)
４．発表年 2019年

１．発表者名 外處 尭之, 瀬戸 謙修
２．発表標題 依存グラフのスケジュール変換による畳み込み処理向けメモリアクセス最適化
３．学会等名 システムとLSIの設計技術研究会 (IPSJ-SLDM)
４．発表年 2020年

１．発表者名 五十嵐 碧, 瀬戸 謙修
２．発表標題 畳み込みニューラルネットワークにおける重みカーネルの分布に基づいた層ごとの最適なプルーニング
３．学会等名 電気学会全国大会
４．発表年 2020年

１．発表者名
伊澤昇平，外處尭之，瀬戸謙修，立岡真人，西田嘉人

２．発表標題
ループ平坦化によるLLVM/Pollyにおけるループ融合の促進

３．学会等名
システムとLSIの設計技術研究会（IPSJ-SLDM）

４．発表年
2021年

１．発表者名
Kenshu Seto

２．発表標題
Scalar Replacement in the Presence of Multiple Write Accesses for Accelerator Design with High-level Synthesis

３．学会等名
DATE 2021（国際学会）

４．発表年
2021年

１．発表者名
坂部 光，瀬戸 謙修

２．発表標題
高位合成を用いたハードウェア設計における三角ループ向けスカラリプレイス

３．学会等名
システムとLSIの設計技術研究会（IPSJ-SLDM）

４．発表年
2021年

１．発表者名
松岡 尚典，瀬戸 謙修

２．発表標題
モンテカルロ木探索と整数線形計画法の組み合わせによる最適スケジューリング

３．学会等名
システムとLSIの設計技術研究会（IPSJ-SLDM）

４．発表年
2021年

１．発表者名 伊澤 昇平，瀬戸 謙修
２．発表標題 ループ平坦化におけるループ回数の2のべき乗化による回路最適化
３．学会等名 システムとLSIの設計技術研究会（IPSJ-SLDM）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関