様式VI

博士学位論文調査報告書

論 文 題 目

Energy Efficient On-Chip Cache Architectures and Deep Neural Network Accelerators Considering the Cost of Data Movement

(データ移動コストを考慮したエネルギー効率の高いキャッシュアーキテクチャとディープニューラルネットワークアクセラレータ)

申請者氏名　　徐　宏傑

最 終 学 歴　　平成　３１年　３月
京都大学大学院情報学研究科通信情報システム専攻修士課程　修了
令和　　３年　３月
京都大学大学院情報学研究科通信情報システム専攻博士後期課程
研究指導認定見込

学 識 確 認　　令和　　年　　月　　日（論文博士のみ）

論文調査委員　京都大学大学院情報学研究科
（調査委員長）　教　授　小野寺　秀俊

論文調査委員　京都大学大学院情報学研究科
教　授　大木　英司

論文調査委員　京都大学大学院情報学研究科
教　授　佐藤　高史

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	徐　宏傑
論文題目	Energy Efficient On-Chip Cache Architectures and Deep Neural Network Accelerators Considering the Cost of Data Movement （データ移動コストを考慮したエネルギー効率の高いキャッシュアーキテクチャとディープニューラルネットワークアクセラレータ）		

（論文内容の要旨）

This thesis aims to propose energy-efficient cache architectures and DNN accelerators by reducing the cost of data movement in various processing tasks. In data processing, the cost of data movement often dominates the hardware performance. Based on factors that affect the prediction capability of the cost of data movement, this thesis investigates representative types of processing tasks as general-purpose processing and special-purpose processing such as Deep Neural Network (DNN). This thesis discusses the processing properties that affect the cost of data movement in each situation. Considering the cost of the data movement, this thesis designs four energy-efficient hardware to explore the effectiveness of the data processing properties. This thesis consists of six chapters.

Chapter 1 provides an overview of the thesis, including the objectives, issues, and approaches of this study.

Chapter 2 proposes an energy-efficient hybrid cache memory system to exploit the non-linear relationship between the cache miss rate and the cache capacity. As the cache capacity increases from zero, the cache miss rate decreases rapidly due to the property of access locality. The rapid decrease however becomes slow down as the capacity of the cache increases. Exploiting this non-linear relationship between the cache miss rate and the cache capacity, a hybrid 2-level on-chip cache architecture which has an energy-efficient Standard-Cell Memory (SCM) as the first level and an area-efficient SRAM as the second level, is first introduced. This thesis then discusses a method for finding the best mix of SCM and SRAM, which minimizes the energy consumption of the hybrid cache under a cache area constraint.

Chapter 3 presents a dense DNN accelerator and an analytical evaluation model based on properties that determine the cost of data movement in DNN processing. In DNN processing, the data movement between memory and processing elements consumes most of time, energy, and area. At the same time, the number of operations in DNN processing is usually much larger than the scale of hardware resources. The key idea is that a whole processing task can be divided into easy-to-handle small tasks for a fine-grained evaluation of the cost of data movement. In this thesis, two hardware properties are investigated to describe the required number of memory accesses and the memory capacity in each memory hierarchy. Based on the two properties, this thesis proposes a DNN accelerator for large-scale processing with a regular access pattern in DNN and a new set of hardware metrics to analytically estimate energy, throughput, and area from processing dataflows. Based on the analytical evaluation of various processing dataflows and various processing environments, the proposed evaluation metrics can effectively evaluate processing dataflows without a hardware implementation. Designers are able to use the proposed evaluation metrics to significantly reduce hardware design effort.

Chapter 4 introduces a sparse DNN accelerator handling irregularity based on the property that

helps enhance the utilization of processing elements in sparse DNN processing. This thesis also addresses a PE utilization issue in sparse DNN processing and sparse DNN properties. The key idea is to directly infer the number of operations from the number of loaded non-zero input-feature-map (ifmap) pixels and the number of loaded non-zero weights. Focusing on the index-matching property in Deep Neural Network, this thesis aims to decompose sparse Deep Neural Network into easy-to-handle processing tasks to maintain the utilization of processing elements. This thesis proposes an efficient hardware accelerator called MOSDA for an irregular data access pattern in sparse DNN. MOSDA does not require complex control logic in sparse DNN processing and speeds up the processing by skipping zero data. According to the case study of this thesis, MOSDA outperforms state-of-the-art CNN accelerators in sparse Alexnet.

Chapter 5 discusses a memory hierarchy which is energy-efficient for Binary Neural Network (BNN) processing. Based on the proposed memory hierarchy, this thesis presents an energy-efficient BNN accelerator. This thesis further proposes a hybrid memory system with an energy-efficient memory architecture for small-scale processing with a regular access pattern in BNN processing. Based on dedicated dataflows for matrix multiplication and convolution, the proposed hardware achieves energy-efficient processing in BNN processing without crossbar logic for accumulating partial sums. The proposed architecture outperforms state-of-the-art BNN accelerators in the case study of this thesis.

Chapter 6 concludes this thesis to summarize processing properties that may affect the cost of data movement and four energy-efficient hardware designs to explore the processing properties.

（論文審査の結果の要旨）

本論文は、データへのアクセスやデータの保存に要するエネルギーならびにハードウェア資源量を削減することによりエネルギー効率を高めたハードウェアの設計法について議論している。データへのアクセスについては、対象とするハードウェアとアプリケーションにより、不規則に発生するものと規則的に発生するものに分類でき、それぞれに応じてアクセスコストの削減を検討する必要がある。本論文では、前者の例として汎用計算機を想定し、階層化構造のキャッシュメモリを導入する事によりエネルギー効率を高める方法を提案した。後者の例としては深層ニューラルネットワークアクセラレータを取り上げ、大規模な密行列演算をおこなう回路、大規模な疎行列演算をおこなう回路、小規模であるが密行列演算を行なう回路のそれぞれについて効果的な実現法を明かにした。本論文で得られた成果は以下の通りである。

1. 汎用プロセッサでは、データのアクセスに規則性はないが局在性があるため、回路面積が小さく高速動作するSRAMを用いたキャッシュメモリが用いられている。本論文では、SRAMキャッシュメモリとCPUの間に、消費エネルギーの小さいSCM(Standard Cell Memory)を配置する階層化ハイブリッドキャッシュアーキテクチャを用いることにより、消費エネルギーが最大68%削減できることを示した。

2. 大規模な密行列の積和演算を実行する深層ニューラルネットワークアクセラレータを、オンチップバッファとレジスタの2段階メモリ階層を持つハードウェアとして実現した。全体処理を階層毎に分割したサブタスクに分解してデータの局所的再利用度を定量化し、データのアクセスコストを削減した並列処理データフローを提案した。Alexnet を実装した場合、最先端のアクセラレータと比較して、消費エネルギーと処理時間が20%削減されることを示した。

3. スパースニューラルネットアクセラレータでは、大規模疎行列の積和演算を実行する。乗算演算の並列処理において、乗数と被乗数の全ての組み合わせを計算するデータフローを用いることにより、不規則に出現する非零要素の演算が効率化できる事を明かにした。提案データフローを用いたアクセラレータでAlexnet を実行した場合、既存のアクセラレータより短い時間で処理を行ない、かつ消費エネルギーを半分以下に削減できる事を示した。

4. 入力や重みを1ビットで表現するバイナリニューラルネットのアクセラレータでは、全てのデータをチップ上に保存する事ができる。オンチップ上でのデータアクセスコストを削減するために、ネットワークデータを保存するSRAMバッファと並列演算部の間にSCMバッファを挿入した階層化ハードウェアを提案した。提案構造を持つハードウェアは、既存のアクセラレータと比較して、エネルギーあたりのデータ処理量が1.4倍になることを示した。

以上、本論文は、エネルギー効率の高いキャッシュメモリや深層ニューラルネットワークアクセラレータの集積回路を実現するための諸問題に関し、データアクセス機構の階層化や並列処理データフローの最適化という観点から解決方法を提案するとともに、回路構成実験やゲートレベル設計回路の評価を行いその有効性を実証している。本論文の内容は、学術上、応用上ともに寄与するところが少なくない。

よって本論文は博士（情報学）の学位論文として価値あるものとして認める。また、令和３年２月１２日に実施した論文内容とそれに関連した試問の結果、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降