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　３ 年　 ５ 月　 ２ 日現在

機関番号：３２６８９
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３２１７

研究課題名（和文）再構成アクセラレータにおけるデータ形式最適化と精度保証

研究課題名（英文）Optimum Data Representation and Its Accuracy Assurance for Reconfigurable Accelerators

研究代表者

木村　晋二（Kimura, Shinji）

早稲田大学・理工学術院（情報生産システム研究科・センター）・教授

研究者番号：２０１８３３０３

交付決定額（研究期間全体）：（直接経費）　　13,500,000円

研究成果の概要（和文）：再構成アクセラレータにおけるデータ形式最適化と精度保証という題目で、FPGA (Field Programmable Gate Array) に代表されるような、ハードウェアの再構成で応用毎に専用のアクセラレータハードウェアを構築できる再構成アクセラレータに対して、データの表現方法と最終的な解の精度を保証する手法の研究を行った。画像処理や画像認識に対して、応用が持つ計算誤差への耐性に基づき、最適なデータ表現および最適な演算器を構築する手法の研究を行った。具体的には、データ表現や演算器に対する誤差の解析手法の提案と評価を行った。

研究成果の学術的意義や社会的意義
近年、CNN (Convolutional Neural Network, 畳み込みニューラルネットワーク)のように、非常に多くの演算を必要とする応用が用いられるようになってきた。そのハードウェアによる高速化は実応用においては非常に重要であり、端末側からサーバー側まで広くハードウェアアクセラレータが用いられている。再構成アクセラレータはそのような応用志向のハードウェアを実現するプラットフォームであり、本プロジェクトで、実際に再構成アクセラレータへ向けたデータ表現とその誤差評価の手法や演算器の提案を行ったことは、学術上および実用上の意義が高い。

研究成果の概要（英文）：The project is on the optimum data representation and its accuracy assurance for reconfigurable accelerators including reconfigurable hardware modules such as FPGA (Field Programmable Logic Array). A reconfigurable accelerator can construct dedicated special hardware accelerators depending on applications. In the optimization of data representation for reconfigurable accelerators, the area, delay and power are optimized under the error tolerance of applications. On image processing and image recognition applications, new data representation methods, operational units for the data representation, and their evaluation methods have been devised and evaluated.

研究分野：ハードウェアの高位設計と検証

キーワード：エラー耐性に基づく最適化　誤差解析　Approximate Computing

１．研究開始当初の背景

近年、画像コーデックや画像認識などの大量のデータを扱いかつ大きな計算時間を要する応用のために、FPGA (Field Programmable Gate Array) など再構成可能ハードウェアを用いたアクセラレータの研究開発が盛んである。具体的には、4K や 8K 向けの動画像のコーデック、CNN (Convolutional Neural Network) を用いた画像認識や、IoT 関連のビッグデータ処理のためのアクセラレータの研究開発が進められている([1], [2])。その裏には、Xilinx による ARM と FPGA の統合チップの販売や、インテルによる Altera の買収と FPGA の CPU への搭載など、CPU と FPGA の融合が進み、マイクロソフトなどのデータセンターでの FPGA の活用が活発化しているなどの事情がある。

例えば CNN に関しては、計算における係数や途中結果の表現において、種々のデータ形式を用いた効率化が研究され、元々の浮動小数点数に代わりに、固定小数点やカスタム浮動小数点を用いる手法、1 と-1 の二値を用いる Binarized CNN や 1, 0, -1 の三値を用いる Ternarized CNN などが提案されている。実際、応用毎にこれらの多様なデータ形式あるいはまだ考えられていないデータ形式を含めて、内部データのデータ形式の最適化を行う必要がある。現状では最適と考えられるデータ形式を試行錯誤で求めざるをえないが、システマティックな最適化手法が求められている。

[1] Eric Chung, “Accelerating Deep Convolutional Neural Networks Using Specialized Hardware in the Datacenter,” The Fourth Workshop on the Intersections of Computer Architecture and Recongurable Logic (CARL 2015), 14 June 2015. http://www.ece.cmu.edu/calcm/carl/ （2017.10.23 Access）

[2] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, Jason Cong, “Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks,” Proc. FPGA15, pp.161-170, Feb. 2015.

２．研究の目的

本研究では、画像処理や画像認識などの誤差を許容する応用のため、再構成可能アクセラレータの速度と電力の効率化を目的として、データの蓄積、転送、処理に適した組込み圧縮を含むデータ形式最適化手法と、データ形式に対する誤差と精度保証の理論およびハードウェア上での評価手法の研究を行う。再構成可能アクセラレータの速度と電力の効率化のためには、応用毎にハードウェアに適したアルゴリズムとデータ形式が必要である。データ表現のビット数の削減は速度および電力の削減に寄与するが、一方で結果の精度に影響を与えるため、精度の低下を抑えながら、データ形式を効率化してビット数を削減することが求められる。誤差と精度の理論的解析の研究を行うとともに、浮動小数点等での評価結果の得られない場合にも適用できる精度解析手法の研究を行う。また、処理途中のデータ形式の動的変更を含む適応的な新データ形式についても研究を行う。

３．研究の方法

まず再構成アクセラレータにおけるデータ形式最適化のために、応用を決めてデータ形式の最適化の研究を行う。具体的には、HEVC (High Efficiency Video Coding) のエンコーダやデコーダおよび CNN のアルゴリズムについて、種々のデータ形式を適用し、最適化の手法を検討する。これまでに、HEVC については、Reconstructed Frame の DRAM との転送・蓄積量の削減に対し、組込み符号化手法の検討を行っており、蓄積・転送のデータ量を大幅に削減する手法を提案している。本手法は、圧縮のための処理量とデータの蓄積・転送量とのバランスをとることに特徴がある。ここではそれを発展させ、種々のデータ圧縮形式を適用した場合の精度についての研究を行う。

圧縮については誤差のないものだけでなく圧縮誤差を伴うものを適用し、その時の誤差解析および精度の評価方法について考える。とくに、品質を一定に保ったままデータのビット幅を動的に変更する手法とその一般化について研究を行う。

CNN に関しては、Ubuntu で caffe に基づく環境を用い、浮動小数点データに対する種々のデータ形式の適用と誤差解析手法の研究を行う。これまでに浮動小数点数のカスタマイズについて、指数部や仮数部のビット数を削減した場合の認識精度の評価を行い、畳込み部の乗算の入力の指数部のビット数を 5 に、また仮数部のビット数を 3 に削減しても十分な認識精度が得られるという結果を得ている(業績 1)。ここではそれを発展させ、畳込みの加算部も含めたカスタム浮動小数点数のビット数の最適化とその時の誤差の解析手法の研究を行う。さらに、浮動小数点や固定小数点にとらわれない動的な適応性を含む新たなデータ形式の検討も行う。

つぎに、これらの応用に対して、データ形式を決めた場合の誤差の伝播と蓄積の計算法について検討を行い、そのハードウェアによる実現を行う。具体的には、積和演算に着目して演算誤差を見積もる手法を研究する。元のデータ表現での値との違いが必要となるが、元のデータ表現での計算を行わずに済ませる手法を開発する。誤差評価のハードウェア化については、アクセラレータ本体の計算構造に付随する形で誤差を伝播させ、それを利用することを考えている。これまでの浮動小数点数の固定小数点化での誤差伝播手法の成果をハードウェア化に応用する。

４．研究成果

平成 30 年度は、画像コーデックにおける中間データの圧縮方法および CNN (Convolutional Neural Network、畳み込みニューラルネットワーク) のデータ形式の最適化に関する研究を行った。とくに、画像データの組込み圧縮法と、畳み込み演算における Approximate Computing 手法の研究を行った。環境整備としては、CNN の開発環境の caffe での誤差解析のための環境整備ならびにその上での誤差の評価実験と、FPGA の高位合成環境の整備とテストを行った。また、データ形式を決めた場合の誤差の伝播と蓄積について計算する手法についての文献調査や理論的な検討を行った。

まず、画像処理のための画像データの組込み圧縮法に関して、情報ロスを含む Lossy 圧縮法の提案を行った。本手法は、量子化と可変長符号化を組合せた方式に基づいており、小さな追加回路で、画像処理回路とメモリの間でデータの送受信を行う際のメモリバンド幅を大幅に削減することを可能とした。

つぎに、CNN における計算の途中結果の誤差を許容できる性質に基づき、演算の一部を簡略化して回路の面積、遅延、電力を削減する Approximate Computing 手法の研究を行い、乗算回路の新たな Approximate 手法をいくつか提案した。乗算では部分積の加算を繰返すが、下位側を OR で近似計算する手法や、部分積の順序を入れ替えることで精度を保持したまま回路を簡単化する手法の検討を行った。

また、Ubuntu で caffe に基づく環境整備を用い、浮動小数点データに対する種々のデータ形式の適用と、乗算と加算における演算誤差の解析方法および誤差の伝播方法の研究を行った。これまでに、浮動小数点数のカスタマイズについて、仮数部のビット数を 3 に、また指数部のビット数を 5 に削減してもトレーニングを含め十分な認識精度が得られるという結果を得ている。ここではそれを発展させ、動的なデータ形式の変更に関する検討を行った。

平成 31 年度／令和元年度は、データの表現形式と精度の関係について理論的な研究を継続して行い、とくに、CNN を題材として、畳み込みにおける高次の Winograd 法で乗算回数を削減した場合の積和演算の誤差や、プーリングなどの非線形関数における誤差、Normalization や学習時の逆方向の演算における誤差の蓄積に関する研究を行った。

ハードウェアによる高速化については、再構成アクセラレータ上で誤差解析を行うハードウェアモジュールの研究を行った。具体的には、現在のデータ形式のハードウェアに少量の回路を付加する方式の検討を行い、指数部や仮数部のビット数を数ビット多めに持つこと、および誤差をデータと同じ精度で持つ手法の研究を行った。

また再構成アクセラレータの効率最適化として、CNN 等の応用で多用される乗算器への Approximate Computing の適用と平均相対誤差距離の評価を行った。整数乗算器の部分積のビット単位での近似法と部分積のグルーピングに基づく近似法の研究を行い、平均相対誤差距離を数 % で抑えつつハードウェアを削減する手法を示した。また、FPGA の LUT (Look Up Table) レベルでの整数乗算器の近似手法を提案した。さらに、Approximate 浮動小数点乗算の研究を行い、手法の提案と同時に解析的な最大誤差評価の手法を示した。

新しいデータ形式に関しては、ブロック浮動小数点数を含むカスタム浮動小数点数の研究と、CNN の重みなど疎行列の表現形式の検討を行った。疎行列では非ゼロの要素のインデックス情報を削減するため、Compressed Row Storage (CRS) 法を拡張した手法の研究を行った。

令和 2 年度は、昨年度に続きデータの表現形式と精度の関係についての理論的な研究を継続し、ResNet など層数の大きい CNN に対して、積和演算だけでなく、もとの値を加える演算、プーリングや relu 活性化関数など非線形関数での誤差、Normalization や学習時の逆方向の演算における誤差の蓄積に関する研究を行った。誤差解析を再構成アクセラレータ上で行う方式については、現在のデータ形式のハードウェアに少量の誤差解析の回路を付加することで、計算結果の誤差の最大を評価するハードウェアの研究を行った。

また、再構成アクセラレータでの電力効率の最適化のために、Approximate Computing に基づく演算器設計と、その誤差の解析手法の研究を行った。今年度はとくに FPGA（Field Programmable Gate Array） 向けの 8 ビット整数近似乗算回路の構成手法の研究を行った。8 ビット乗算を 4 つの 4 ビット乗算結果の和で表し、4 ビット乗算を FPGA の LUT (Look Up Table) 素子を用いて構成する手法を 3 通り示した。これらは使用する LUT 素子の数と計算精度が異なる。それらの組合せで精度の異なる 8 ビットの乗算回路を構成した。既存のものと比べて同じ計算精度の場合は、消費エネルギーに相当する電力と遅延の積を小さくすることができている。研究成果は学術論文誌に掲載された。また、ASIC 向けに浮動小数点乗算器の近似化に関する研究も行った。浮動小数点乗算では、仮数部の上位 4 ビットに着目し、上位 4 ビットについては正確な乗算を、また残りの部分については非ゼロの 4 ビット部分を求めて、それとの計算結果を近似に用いる方式を提案した、研究成果は国際会議 TENCON 2020 において発表した。

疎行列の表現方法についても研究を継続し、表現方法の実行時間への影響についての検討を行った。また、浮動小数点数の整数化によるデータ量削減の効果と誤差評価についても検討を行った。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　3件／うちオープンアクセス　1件）

１．著者名 Yi GUO, Heming SUN, Ping LEI, and Shinji KIMURA	４．巻 E103.A
２．論文標題 Approximate FPGA-Based Multipliers Using Carry-Inexact Elementary Modules	５．発行年 2020年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 1054-1062
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.2019KEP0002	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 GUO Yi、SUN Heming、LEI Ping、KIMURA Shinji	４．巻 E102.A
２．論文標題 Design of Low-Cost Approximate Multipliers Based on Probability-Driven Inexact Compressors	５．発行年 2019年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 1781～1791
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E102.A.1781	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Li Guo, Dajiang Zhou, Jinjia Zhou, Shinji Kimura, and Satoshi Goto	４．巻 6
２．論文標題 Lossy Compression for Embedded Computer Vision Systems	５．発行年 2018年
３．雑誌名 IEEE　Access	６．最初と最後の頁 39385-39397
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2018.2852809	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 該当する

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　7件）

１．発表者名 Jie Li, Yi Guo, and Shinji Kimura
２．発表標題 Accuracy-Configurable Low-Power Approximate Floating-Point Multiplier Based on Mantissa Bit Segmentation
３．学会等名 IEEE Region 10 Conference（国際学会）
４．発表年 2020年

１．発表者名
Jie LI, Yi GUO, and Shinji KIMURA

２．発表標題
Approximate Floating Point Multiplier based on Shifting Addition Using Carry Signal from Second-Highest-Bit

３．学会等名
IEICE Tech. Report, VLD2019-120

４．発表年
2020年

１．発表者名
Guo Yi、Sun Heming、Kimura Shinji

２．発表標題
Small-Area and Low-Power FPGA-Based Multipliers using Approximate Elementary Modules

３．学会等名
Proc. of ASP-DAC 2020（国際学会）

４．発表年
2020年

１．発表者名
Li Guo, Dajiang Zhou, Jinjia Zhou, Shinji Kimura

２．発表標題
Embedded Frame Compression for Energy-Efficient Computer Vision Systems

３．学会等名
ISCAS 2018（国際学会）

４．発表年
2018年

１．発表者名
Li Guo, Dajiang Zhou, Jinjia Zhou, Shinji Kimura

２．発表標題
Sparseness Ratio Allocation and Neuron Re-pruning for Neural Networks Compression

３．学会等名
ISCAS 2018（国際学会）

４．発表年
2018年

１．発表者名
Yi Guo, Heming Sun, Li Guo, Shinji Kimura

２．発表標題
Low Cost Approximate Multiplier Design using Probability Driven Inexac t Compressors

３．学会等名
APCCAS 2018（国際学会）

４．発表年
2018年

１．発表者名
Zhenhao Liu, Yi Guo, Xiaoting Sun and Shinji Kimura

２．発表標題
Energy-Efficient and High Performance Approximate Multiplier Using Compre ssors Based on Input Reordering

３．学会等名
TENCON 2018（国際学会）

４．発表年
2018年

１．発表者名
Xiaoting Sun, Yi Guo, Zhenhao Liu, Shinji Kimura

２．発表標題
A Radix-4 Partial Product Generation-Based Approximate Multiplier for High-Speed and Low-Power Digital Signal Processing

３．学会等名
ICECS 2018（国際学会）

４．発表年
2018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
研究分担者	戸川　望 (Togawa Nozomu) (30298161)	早稲田大学・理工学術院・教授 (32689)	

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