学位被授与者氏名	席　家禎（XI JIAZHEN）
学位の名称	博士（工学）
学位番号	博（一）第６５号
学位授与年月日	２０２１年9月１６日
論文題目	Improving Bit Energy Efficiency of Machine Learning Systems Towards an Era of AI-Everywhere
論文題目 （英訳または和訳）	どこでも AI に向けた機械学習システムのビットエネルギー効率向上
論文審査委員	論文審査委員会 委員主査　：福岡工業大学大学院知能情報システム工学専攻教授　山内　寛行 同審査委員：福岡工業大学大学院知能情報システム工学専攻教授　種田　和正 同審査委員：福岡工業大学大学院知能情報システム工学専攻教授　福本　誠 同審査委員：福岡工業大学大学院物質生産システム工学専攻教授　江口　啓
論文審査機関	福岡工業大学大学院工学研究科
論文内容の要旨 （和文）	近年のディープラーニング型ニューラルネットワーク(NN)の機械学習システムはより複雑な画像の認識精度を畳み込み層をより深くして向上させている。結果、パラメータ数や乗加算計算回数、計算対象のデータの総移動距離/時間の増加に伴う急激な消費電力増加の課題が顕在化している。さらに今後、クラウドだけでなくセンサー付近にも AI を埋め込む AI-IoT（どこでも AI）のシステムの実現を考えた場合、今とは桁違いにエネルギー効率の高い機械学習システムの実現なくしてはその夢はかなえられない。本論文ではその目標の深刻な障害となる課題を解決する技術の提案と効果、残された課題について、以下の 2 つの提案技術から論ずる： 1） Memory 回路にデータ記憶機能だけでなく乗加算計算（MAC）の機能も Dual Role として兼務させる In-Memory Computing 技術でデータ移動に起因するエネルギー削減手法を論じ、メモリアレーのコラム数削減手法の提案とその効果(16%削減)、課題を論ずる。 さらに、 2） MAC 計算対象データの表現ビット数を従来の 32bit 浮動小数点でなく大胆に 1bit (Binary Neural Network: BNN)に削減することで、Bit 演算のエネルギー効率を桁で向上させることを目的とした低精度表現 NN 機械学習モデルの精度や安定性の問題を解決する手法の提案とその効果(15%精度向上と 54%のばらつき削減 on CIFAR-10 dataset)、と課題を論ずる。 第 1 章では 1bit を含む低精度表現型 NN 機械学習モデル、BNN 機械学習の概念と目的を説明し、従来の BNN が抱える課題を明らかにし本論文の背景と動機を明らかにする。 第 2 章では従来の In-Memory 技術が微細化や低電圧動作に伴う MAC 演算回路特性の非線形性のために生ずるモデル誤差の解決のために導入していたアンサンブル学習の副作用である面積、消費電力課題を解決する手法を提案する。 第 3 章では従来の低精度表現型 NN 機械学習モデルが抱えるビット表現力の低下による精度劣化と不安定な学習状態を解決する手法として、注目する層にのみバギング学習を適用し、BNN での表現力低下を抑制し、従来の BNN 学習において課題であったトレーニング曲線の大きなふらつきを抑制する手法を提案する。又、その効果を汎用的に評価するために広範囲に異なる NN 機械学習モデルのネットワーク構成に適用した場合、あるいは CIFAR-10 以外のデータセットに適用した場合の詳細な実験結果を取得し、その効果を汎用的に議論する。 第４章では、BNN 学習の不安定性の課題をさらに掘り下げ、いきなり１ビット化せずに段階的に徐々に 1 ビット化するなどの緩和プロセスを提案し、別の視点から提案されているバッチ正規化プロセスや Adam などのオプティマイザーに対する依存性などを明らかにしながら提案技術の効果を汎用的に議論する。 第5章では本論文の主要な結果を要約する。
論文内容の要旨 （英文）	The power consuming amount of neural network has intolerably increased with ever-increasing amount of number of MAC (Multiply Accumulate) operations and data movements

	for computing the machine learning models. Thus, the computing performance for the models can't be increased any more under the power constraint systems until much higher energy efficient model could be developed. To address such challenge, some trials with using in-memory computing and binary neural networks (BNN), has drawn much attention in the power-constraint fields like the internet of things (IoT). However, the 1-Bit training process of conventional methods also needs much training memory/time at a cost of the unstable training process and the loss of accuracy. This work proposed (1) A column reduction technique for in-memory machine learning classifier. The proposed method can achieve the similar accuracy as original full precision with MNIST dataset and with much lower computing memory (with 16% columns reduced), compared to conventional work. (2) Layer-wise ensemble technique for BNN to improve the performance of low precision networks via employing the ensemble learning technique which reduces the error and its standard deviation by 15% and 54% on CIFAR-10 dataset, respectively, compared to the BNN serving as a baseline. (3) Training with relaxation of both weights and activations for binary neural networks to alleviate the variance by the conventional BNN training process (reduced by 2%) with the experiments of various cases including different optimizers and with or without batch normalization. The outline of this thesis is as follows. Chapter 1 presents the concept of low precision machine learning and the purpose of this work. Chapter 2 introduces the conventional in-memory boosting classifier and then proposes a column reduction technique to improve the bit energy efficiency of the in-memory boosting classifier followed by the comparison to the conventional technique. Chapter 3 presents the low precision neural network and its instability issue of conventional works and then proposes a layer-wise ensemble method for the binary neural network (BNN), followed by the experiments to compare the conventional methods and the proposed one the metrics of the stability and accuracy of the network. Chapter 4 presents the conventional training process of BNN with relaxation technique and proposes a new training procedure which employs relaxations to both low precision weights and activations, followed by the comparisons between proposed method and conventional training process in various cases of the different optimizers and with/without batch normalization. Chapter 5 summarizes this thesis.
論文審査結果	博士後期課程知能情報システム工学専攻 3 年のセキ　カテイ氏が提出した博士学位論文を審査し、また、最終試験を実施したのでその結果について報告する。 **〈学位論文審査結果〉** 近年のディープラーニング型ニューラルネットワーク(NN)の機械学習システムはより複雑な画像の認識精度を畳み込み層をより深くして向上させている。結果、パラメータ数や乗加算計算回数、計算対象のデータの総移動距離/時間の増加に伴う急激な消費電力増加の課題が顕在化している。一方で、今後、クラウドだけでなくセンサー付近にも AI を埋め込む AI-IoT（どこでも AI）のシステムの実現を考えた場合、今とは桁違いにエネルギー効率の高い機械学習システムの実現なくしてはその夢はかなえられない。 本論文ではその目標達成に向けた課題を解決する技術の提案と効果、残された課題について論じている。 最初に、Memory 回路にデータ記憶機能だけでなく乗加算計算（MAC）の機能も Dual Role として兼務させる In-Memory Computing 技術でデータ移動に起因するエネルギー削減手法を論じ、メモリアレーのコラム数削減手法の提案とその効果(16%削減)を示している。 さらに、MAC 計算対象データの表現ビット数を従来の 32bit 浮動小数点でなく大胆に 1bit (Binary Neural Network: BNN)に削減し bit 演算のエネルギー効率を桁で向上させた。その副作用として出現する機械学習モデルの精度や安定性の課題を解決する手法として、注目する層にのみバギング学習を適用し、BNN での表現力低下を抑制し、従来の BNN 学習において課題であったトレーニング曲線の大きなふらつきを抑制する手法を提案している。その結果、15%精度向上と 54%のばらつき削減(on CIFAR-10 dataset)を達成している。

	BNN 学習の不安定性の課題もさらに掘り下げ、いきなり 1 bit 化せずに段階的に徐々に 1bit 化するなどの緩和プロセスとその対象を重みパラメータだけではなくアクティベーション層にも適用し、重みパラメータだけに適用した場合と比較して 1.6%の精度向上と 2.5%のばらつき削減(on CIFAR-10 dataset)を達成している。 本論文は 5 章から構成されている。本論文中には、研究の背景、目的、提案技術、工学的意義、独創性等を明確に記述している。また、当該分野における将来の研究課題や展望の見識が的確に述べられている。また、本研究の一部を学術論文としてまとめ、2 編（第一著者 2 編）発表済みである。 以上の点などを考慮した学位論文評価ルーブリック I に基づく全審査委員の評価において学位論文として評価できるレベルに達していると認められた。 〈**最終試験報告書**〉 令和 3 年 8 月 5 日の学位論文公聴会においては、論文内容に関連する種々の工学的及び技術的な質問があったが、いずれも適切な回答を行うことができた。最終試験においては、公聴会後の学位論文評価ルーブリック II に基づく全審査委員の評価において、学位論文として評価できるレベルに達していると認められた。 以上の結果から、学位審査委員会はこの論文が博士（工学）の学位に適格であると判定した。
主な研究業績	参考論文　6 編 1 冊 査読付き学術論文：第一著者 2 編 1. "A Column Reduction Technique for In-memory Machine Learning Classifier", International Journal of Machine Learning and Computing, Vol. 8, No. 2, DOI: 10.18178/ijmlc.2018.8.2.675, Apr. 2018 Authors: Jiazhen Xi, Hiroyuki Yamauchi 2. "A Layer-wise Ensemble Technique for Binary Neural Network", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 35, No. 8, DOI: 10.1142/S021800142152011X (in press), Jun. 2021 Authors: Jiazhen Xi, Hiroyuki Yamauchi 査読付き国際会議論文：1 編 1. "Time and Environment Dependency Aware Fuel Consumption Tracking Method for Improving Drivers and Trucks Management", Circuits / Systems, Computers and Communications (ITC-CSCC), The 36th International Technical Conference on, ITC-CSCC 2021, Jun. 2021 Authors：Yu Peng, Xi Jiazhen, Hiroyuki Yamauchi 査読なし国際会議論文：第一著者 1 編 1. "Layer-wise Ensemble for Binary Neural Networks", The 1st NKUST-FIT International Seminar on Advanced Technology, Dec. 2018 Authors：Jiazhen Xi and Hiroyuki Yamauchi 査読なし国内会議論文：第一著者 1 編 1. "Approximately Quantizing Algorithm for In-memory Machine Learning Classifier", The 17th Forum on Information Technology, Sep. 2018 Authors：Jiazhen Xi, Tsuru Ryusuke, Hiroyuki Yamauchi