ALCA-Next
「グリーンコンピューティング・DX」領域
2024 年度　年次報告書

2024 年度採択

［研究開発代表者名：本村　真人］

［東京科学大学総合研究院　教授］

［研究開発課題名：グリーンで信頼される AI を支えるシリコンブレインキューブの実現］

主たる共同研究者：
［安藤　洸太（北海道大学情報科学研究院　准教授）］
［安戸　僚汰（京都大学情報学研究科　助教）］

実施期間　：　2024 年 11 月 1 日～2025 年 3 月 31 日

§１．研究開発成果の概要

11 月の研究開始後、Imperial College London との共同研究プラットフォーム(Slack, ホームページ、隔週ミーティング等)を立ち上げ、計画書 WP1-6 で定義した具体的な研究活動を開始した。

(WP1)宝くじ仮説に基づく LLM(Large Language Model)小型化について、国際会議 NeurIPS Workshop で 2 件の論文を発表し、TMLR でジャーナル論文発表した 1)。また Transformer アーキテクチャの計算量・メモリアクセス削減を目指し、第一ステップとして ViT に焦点を当て、トークン選択からマージ方法に至るまで全プロセスを見直した包括的な動的トークン削減アプローチを提案した。更にエネルギー関数が二次式に制約されない拡張イジングモデルに対するアニーリング手法を構築し、Swin Transformer の事後トレーニング量子化（PTQ）に活用した。ファイチューニングなしで高精度な PTQ が可能であることを確認し、現在、本成果に基づいて論文を執筆中である。

(WP2) Processing-in-Memory（PIM）における LLM のロングシーケンスタスクの推論時の課題が、活性値のメモリ間データ移動コストにあること、また、既存のアテンション軽量化手法の PIM への応用が難しいことにあることを発見し、新たに PIM を前提とした活性値圧縮手法を提案し性能評価を行った。LLM・Transformer 推論の軽量化に向け、固定構造の学習済み部分を動的に選択して組み合わる手法を検討し、予備評価を進めた。

(WP3) 先行して進めていた不規則スパースなモデルを効率よく実行可能なハードウェア技術(Triple Unstructured Sparsity Exploitation)に基づく LSI の DNN トレーニング技術、アーキテクチャ、回路技術、評価結果を国際会議 A-SSCC で発表した 2)。

(WP5) 英国側と密に連携して研究チームを立ち上げ、英国側が持つ設計最適化ツールと信頼性のある AI の技術と日本側が持つエネルギー効率の良いハードウェア志向機械学習モデルの技術を組み合わせて新たな価値を生むことを目的として研究を進めた。本年度は第一ステップとして、ベイジアンニューラルネットワークモデルを、複数の指標に基づいて自動で最適化する手法を開発した。その成果をまとめた国際共同論文を国際会議 HEART に投稿して採択が決定した 3)。

【代表的な原著論文情報】

1) Hikari Otsuka, Daiki Chijiwa, Ángel López García-Arias, Yasuyuki Okoshi, Kazushi Kawamura, Thiem Van Chu, Daichi Fujiki, Susumu Takeuchi, Masato Motomura, “Partially Frozen Random Networks Contain Compact Strong Lottery Ticket,” Transactions on Machine Learning Research (TMLR), Feb. 20, 2025.
2) Yasuyuki Okoshi, Ángel López García-Arias, Jaehoon Yu, Junnnosuke Suzuki, Hikari Otsuka, Thiem Van Chu, Kazushi Kawamura, Daichi Fujiki, Masato Motomura, “WhiteDwarf: 12.24 TFLOPS/W 40 nm Versatile Neural Inference Engine for Ultra-Compact Execution of CNNs and MLPs Through Triple Unstructured Sparsity Exploitation and Triple Model Compression,” 2024 IEEE Asian Solid-State Circuits Conference (A-SSCC), Nov. 21, 2024.
3) Zhiqiang Que, Hongxiang Fan, Gabriel Figueiredo, Ce Guo, Wayne Luk, Ryota Yasudo, Masato Motomura, “Trustworthy Deep Learning Acceleration with Customizable Design Flow Automation,” International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART), May 26, 2025 (Accepted)

以上