

2024 年度年次報告書
信頼される AI の基盤技術
2022 年度採択研究代表者

富岡 洋一

会津大学 コンピュータ理工学部
上級准教授

持続可能な高効率 AI システムの実現

研究成果の概要

ミッションクリティカルシステムにおいては、経年劣化などに起因する AI 回路の故障が深刻な誤動作を引き起こす可能性がある。そこで本研究では、故障の正確な検出および故障後も高精度な認識を継続可能な耐故障 AI を、低計算量・小面積で実現する技術の確立を目的とする。

本年度は低コスト、低消費電力の耐故障 AI 実現のために、(1) 残差量子化に基づく近似 Triple Modular Redundancy (TMR)、(2) アンサンブル耐故障 CNN のための Federated Deep Mutual Learning (FDML)、(3) 低ビット量子化・アンサンブル・Dual Modular Redundancy (DMR) の統合に取り組んだ。

(1) 近似 TMR は、消費電力の低減を目的として、CNN 各層においてオリジナルモジュールと、それを低ビットに量子化した 2 つのモジュールを組み合わせで構成される。低ビット量子化モジュールとの出力誤差により、故障の発生した部位を特定することができる。さらに、オリジナルモジュールが故障した際には、低ビットモジュールの出力を統合することで代替機能を果たし、出力精度の劣化を抑制できる。ただし、従来の線形量子化では、低ビット出力が元の出力と大きく乖離し、故障時に大幅な精度低下を引き起こす問題があった。これを解決するために残差量子化を導入し、故障した層に対して複数回の低ビット演算を行うことで、故障前と同等の精度を実現した。実際に回路を実装した結果、消費電力を約 40% 削減できることを確認している。さらに、Vision Transformer モデルに対しても故障耐性を解析し、故障の影響を受けやすい層の特定と、近似 TMR による高精度な故障検出が可能であることを示した。

(2) 昨年度に提案したアンサンブル耐故障 CNN に対し、Deep Mutual Learning (DML) と Federated Learning (FL) を統合した新たな学習手法 Federated Deep Mutual Learning (FDML) を開発し、精度の向上を実現した。このアンサンブル耐故障 CNN では、各子モデルを個別のアクセラレータで実行し、出力結果の共通性に基づいて故障を検出する。同時に、正常な子モデルの出力を平均化することで、より大規模なモデルに匹敵する精度を達成する。FDML では、「DML による相互学習」と「一部計算の共有化」を交互に行うことで、計算の共通性を徐々に高める耐故障アンサンブルへと誘導する。学習後のモデルに対してソフトウェアの耐性評価を実施し、従来の単一モデルと比較して大幅な耐故障性向上が確認された。

(3) さらに、低ビット量子化、アンサンブル、DMR を組み合わせることで、計算コストの削減と耐故障性の両立を図った新しい手法を提案した。この手法により、高い信頼性を有する路面リスク検出モデルを構築し、その有効性を確認した。

【代表的な原著論文情報】

- 1) 2) Y. Saikawa and Y. Tomioka, "Approximated Triple Modular Redundancy of Convolutional Neural Networks Based on Residual Quantization," *2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, Kuala Lumpur, Malaysia, 2024, pp. 302-309, doi: 10.1109/MCSoc64144.2024.00057.
- 2) M. Ishii, K. Su, Y. Tomioka and H. Saito, "Efficient and Fault-tolerant Object Localization and Classification Based on an Ensemble of Dual Ternary YOLIC Models," *2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, Kuala Lumpur, Malaysia, 2024, pp. 310-315, doi: 10.1109/MCSoc64144.2024.00058.

Lumpur, Malaysia, 2024, pp. 332-338, doi: 10.1109/MCSoC64144.2024.00061.