2023 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では、経年劣化等による 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、故障を正確に検知し、故障後も良好な認識を継続できる耐故障 AI を低計算量、小面積で実現する技術を確立することを目的としている。

従来の耐故障技術である Dual Modular Redundancy (DMR)は同一の回路を2個並べ、それらの計算結果を比較することで故障を検出する。一方、昨年度提案した近似 DMR は、CNN の各層とより低ビットに量子化した層を組み合わせ、それら出力の平均絶対誤差が閾値以上の場合に故障を検出する。ResNet-20 の近似 DMR 回路を設計し、リソース使用量と消費電力を評価し、近似 DMR により従来 DMR に対して Look-Up Table (LUT)使用量を約 27%、消費電力を約 19%削減できることを確認した。また、Cifar-10 データセットで学習した ResNet-20 に対して近似 DMR を用いることで2層を除いて約 1.0 の Area Under Curve (AUC)を達成できることを確認した。

エッジ AI アクセラレータを3つ並べて同じ推論を実行し、その出力の多数決をとる Triple Modular Redundancy (TMR)を利用することで、ひとつのエッジ AI アクセラレータが故障した場合にも高精度な推論を継続することが可能となるが、消費電力が3倍に増加する問題がある。そこで本研究では、より低消費電力の小型エッジ AI アクセラレータを複数組み合わせて、アクセラレータの故障検出と高精度かつ高速な推論を同時に達成する新しいアンサンブル耐故障 CNN を提案した。アンサンブル耐故障 CNN では、各子供モデルを別のアクセラレータで実行し、子供モデルの共通する計算の出力により故障を検出しつつ、正常な子供モデルの出力の平均により、より大型なモデルと同程度の精度を達成する。異なる教師モデルからの知識の蒸留を用いることで子供モデルの多様性を向上させ、従来の TMR に比べて半分以下の計算量で、同程度の推論精度を達成した。