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

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。このため、AI に突発的な故障が発生した場合にも、故障を検出し適切な出力を維持できる耐故障 AI を実現していくことが必要不可欠である。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。

第一年次は、畳み込みニューラルネットワーク（CNN）の耐故障化性能評価、近似に基づく小面積の CNN 故障検出に取り組んだ。最初に、経年劣化による遅延の増加、ソフトエラーによるパラメータのビット反転を想定し、畳み込み層の重み、バイアス、出力特徴マップの各要素に対して一定確率で故障を発生させる故障畳み込み層を実装し、故障が推論精度に及ぼす影響を調査した。実験では、検査対象モデルの各畳み込み層を故障畳み込み層に置換し、評価用データを用いて推論精度の低下を評価した。この結果、連続する畳み込み層では入力側の層の方が故障耐性は低いといった CNN の故障耐性の傾向を明らかにした。

既存の耐故障化技術として Dual Modular Redundancy (DMR)がある。DMR では、同一の回路を2個並べて、それらの計算結果を比較することで故障を検出する。CNN のアクセラレータの故障検知においても DMR を用いることができるが、この場合、アクセラレータに必要な回路資源、推論に要する演算量が 2 倍になる。この結果、推論に要するエネルギーも増加する。そこで、8 ビット量子化 CNN の DMR における一方のアクセラレータを近似計算に基づき小型化し、元の計算結果と近似計算の結果を比較することで故障を検出する近似 DMR 技術を提案した。ResNet-20 を用いた実験では、ほとんどの層において推論精度が大きく低下する前に高精度に故障を検出することが可能であることを確認した。