FPGA 実装 AI アクセラレータにおけるアーキテクチャと推論時ソフトエラー耐性の評価

鉄井 誠人（量子･古典集積回路研究室）
（指導教員 廖 望 講師）

1. はじめに

機械学習の普及により、組込み向けに小型かつ高性能な並列計算デバイスである FPGA（再構成可能な回路）を用いた AI（人工知能）アクセラレータの実装が注目されている。一方で、放射線起因の一時的故障であるソフトエラーは、回路構成情報を保持する CRAM（構成メモリ） のビット反転によって誤動作を引き起こす可能性がある[1]。

本研究では、回路規模および処理サイクル数の異なるアーキテクチャの AI アクセラレータを FPGA 上に実装し、推論時にソフトエラー注入体制を構築した。これにより、アーキテクチャの違いが故障率および故障発生時の推論結果に与える影響を評価し、耐エラー性の高い設計指針の獲得を目的とする。

2. アクセラレータのアーキテクチャ設計

アクセラレータのアーキテクチャ設計として、(1) 16 回の積和演算(MAC)を並列に実行可能な PE（プロセッシングエレメント）を 16 個配置する構成（16-16MAC）、(2) PE の配置数を保ちながら、各 PE における積和演算の並列度を減らし、積和演算を時間分割で順次実行する構成(16-1MAC)を設計した。両構成は同一の演算内容を処理し、回路規模および処理サイクル数の異なるアーキテクチャで比較できるようにした。

演算モデルとして、ニューラルネットワークを固定小数点に量子化し、同じモデルと入力に対して、エラーによって生じるアクセラレータのハードウェア誤動作について評価する。以降はハードウェア誤動作を故障と定義する。

3. アクセラレータのソフトエラー評価手法

FPGA に構築したアクセラレータのソフトエラー評価は、低コストで実施可能なエラー注入手法を用いた。アクセラレータのアーキテクチャ部分の CRAM のみを注入対象とし、エラー注入を各ビット順次に実施した。

故障は、推論結果への影響度が大きい順に、(1)アクセラレータ機能停止、(2)一部の PE において演算結果が 0 に固定される 0 スタック、(3)複数の PE 出力誤り、(4)1 つの PE 出力誤りの 4 種類に分類して評価した。

エラー注入後のアクセラレータ信頼性比較として、推論 1 回あたりの AVF（構成的脆弱因子）の値を用いる。AVF は、エラー注入によって発生した故障の割合を表す指標であり、その値が小さいほど信頼性は高くなる。

$$AVF = \frac{\sigma \cdot P_{critical} \cdot N_{eb}}{factor_{inf}} \quad (1)$$

$$factor_{inf} = \frac{1}{N_{cycle}} \quad (2)$$

推論結果に現れた乖離は、エラー注入時の推論結果と正常時の推論結果との差を、正常時の結果で正規化した結果乖離度として式 (3) のように定義し、推論結果に与えるエラーの影響を定量的に評価することが可能となる。

$$\text{結果乖離度} = \frac{\left|\text{注入時推論結果} - \text{正常時推論結果}\right|}{\text{正常時推論結果}} \cdot 100 \quad (3)$$

4. 実験結果

エラー注入の結果を表 1 に、その故障の内訳を図 1 に、各アーキテクチャ事の結果乖離度の区間統計を図 2 に示す。故障を起こした CRAM ビットをクリティカルビット(CB)として定義する。

表 1　エラー注入の結果

アーキテクチャ	CRAM 数	CB 数	故障率
16-16MAC	2937276	32188	1.1 %
16-1MAC	168643	15247	9.04 %

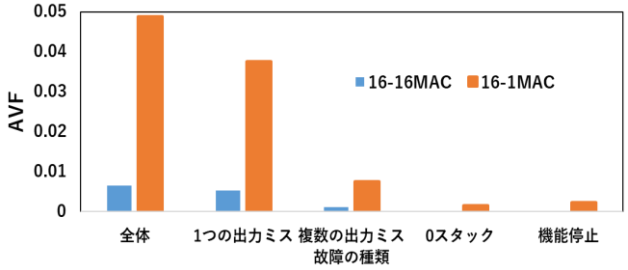

図 1　信頼性評価の結果

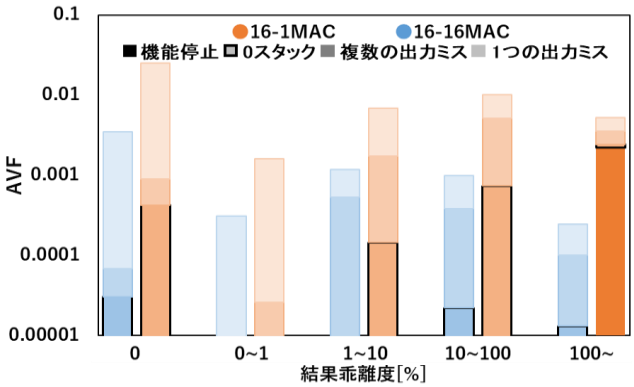

図 2　結果乖離度区間統計

以上の結果より、全体 AVF の観点では、16-16MAC 構成は 16-1MAC と比較して低い値を示しており、推論時ソフトエラーに対して高い信頼性を有することが確認された。一方、16-1MAC では、 0 スタックや機能停止といった影響度の大きい故障の割合が相対的に高く、推論結果へ大きく影響する傾向が見られた。また、CRAM 数が少ないにもかかわらず、推論回数当たりの AVF も高い点は、時間多重による影響集中を示唆している。

また、結果乖離が発生しなかったものを除外して結果乖離度の分布を比較すると、16-16MAC では 1～10%の区間に分類されるケースが最も多く分布しているのに対し、16-1MAC では 10～100%の区間に分類されるケースが最多となった。

これらの結果から、信頼性および故障発生時の影響度の両観点において、16-16MAC 構成が 16-1MAC 構成よりも優れた特性を有することが確認された。

5. まとめ

本研究では, 回路規模および処理サイクル数の異なる AI アクセラレータアーキテクチャを FPGA 上に実装し、推論を行う時にエラー注入を用いた評価体制を構築し，故障率および結果乖離度を指標として、アクセラレータのアーキテクチャとエラー影響の関係を評価した。

参考文献

[1] I. Souvatzoglou et al., “The Impact of Hardware Folding on Dependability in Space-borne FPGA-based Neural Networks,” ICFPT, 2022, pp. 1–1.
[2] I. Souvatzoglou et al., “Assessing the Reliability of FPGA-based Quantized Neural Net-works Under Neutron Irradiation,” IEEE TNS, vol. 71,no. 12, pp. 2565–2577, 2024.