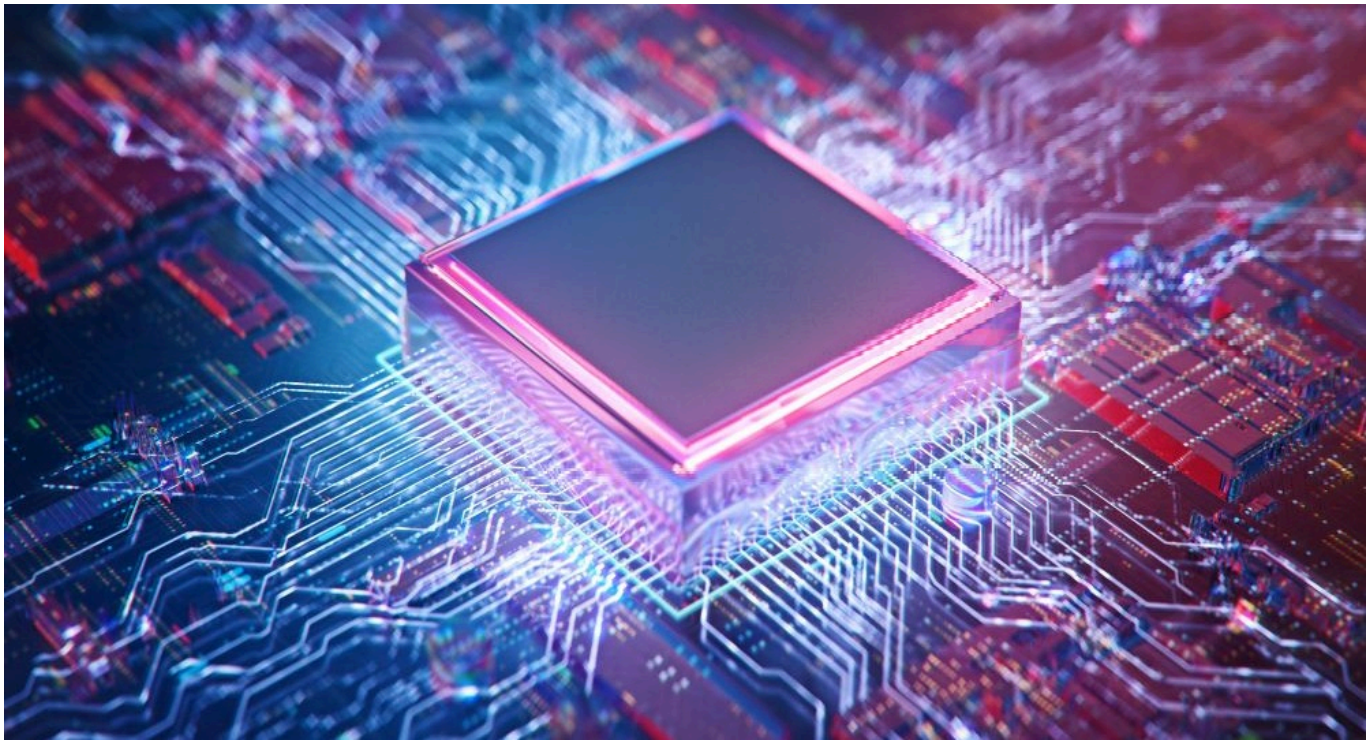
TECH OFFER

Fault-Tolerant Technique For Deep Learning Accelerator Design



KEY INFORMATION

TECHNOLOGY CATEGORY:
Electronics - Embedded Systems

TECHNOLOGY READINESS LEVEL (TRL): TRL4
COUNTRY: SINGAPORE
ID NUMBER: TO174410

OVERVIEW

Successful fault-injection attacks on Deep Neural Network (DNN) hardware accelerators for real-world object recognition and classification problems have been demonstrated. For safety and security-critical applications, fault resiliency on computation errors is a desirable property to avert DNN misclassification. Mitigation of computation errors require real-time error detection and correction. Error correction codes are not applicable to arithmetic operations; while doubling or triplicating the functional units with majority voting to achieve fault tolerance is too expensive. Existing lightweight shadow register-based error correction solutions need to stall the execution upon error detection to flush the multiply-accumulate (MAC) pipeline. To recover from the fault, the data will have to be replayed and re-computed. This process introduces non-trivial throughput and power consumption overheads. Other fault resilient DNN hardware designs exploit the inherent redundancy of the DNN model to avoid explicit error correction. However, such design methodologies can only recover the prediction accuracy from uniform errors or sparse faults due to natural noises or particle radiation but not biased and intensive faults arising from deliberate attacks.

This technology offer is a hardware design method for DNN convolution operations that provides efficient and timely error

correction to prevent prediction accuracy degradation due to both naturally occurring errors as well as maliciously injected faults. Specifically, this design method can be applied to the convolutional layers of any pre-trained DNN models for efficient implementation on both application-specific integrated circuit (ASIC) and field programmable gate array (FPGA) platforms to increase its robustness against fault-injection attacks without impacting the original throughput.

TECHNOLOGY FEATURES & SPECIFICATIONS

The convergence of Artificial Intelligence (AI) and Internet of Things (IoT) brings new challenges for the security of deployed endpoint devices. Data analytics at the edge is vulnerable to fault injection attacks due to both physical and remote accessibility to internal operations of DNN by an adversary. This solution can be implemented on a low-cost edge computing platform to enable DNN computing errors to be recovered in a timely manner before they are escalated to cause a misclassification or wrong prediction under a malicious fault-injection attack. This method is different from existing data replay or approximation schemes in that it can preserve the throughput and accuracy of pre-trained DNN model under deliberate attacks such as overheating, voltage sagging, clock glitching and overclocking, etc. More importantly, this solution can detect the computing errors in real-time and restore the correct computations without suspending the existing pipeline or requiring data replay. Hence, this implementation incurs no performance penalty and only fractional hardware and negligible power consumption overheads. This method works on any hardware architecture dominated by pipelined MAC operations such as the convolutional layers and batch normalization in tensor processing units and machine learning accelerators. Based on a FPGA prototype evaluation, this technique of MAC design can achieve 12.2% to 47.6% higher error-resiliency than existing fault mitigation methods with the same fault injection rates.

POTENTIAL APPLICATIONS

This technique can lead to a lightweight error-resilient and hardware-efficient implementation of pretrained deep convolutional neural network models. A robust and fault tolerant DNN accelerator is an essential element for edge AI applications, such as:

1. Defect detection on manufacturing process.
2. Traffic management in transportation.
3. Queue detection and in-store automated checkout in retailing.
4. Farmland monitoring in smart agriculture.
5. Entertainment application for virtual reality and augmented reality.

The technology owner is keen to license this hardware design to ASIC or FPGA design houses developing DNN accelerator integrated circuits.

UNIQUE VALUE PROPOSITION

Edge computing platforms with on-device inference capability is a new paradigm for AI deployment, particularly when transmission bandwidth or network reliability is a main concern. Businesses can benefit from a more secure and reliable deep neural network processing on many real-time edge AI use cases. The advantage of our method is that it can be easily adapted to existing MAC pipelines in DNN accelerator design without throughput and prediction accuracy penalty at the cost of a small hardware overhead.