**Fault injections against CNN’s on systolic arrays**

Abstract

CNNs play a major part in image classification today, a method which is widely used in autonomous cars, IoT devices and wide-spread ML algorithms.

The baseline which all these use cases rely on is the accuracy of the trained model, where in the case of autonomous cars, the accuracy of the model has a direct impact on the safety of its usage, a case usually called functional safety (FuSa) violation.

Fault injections when introduced to the datapath of a systolic array may have a heavy impact on the accuracy of the trained model, such that it may induced FuSa violations.

In this paper, we measure the classification error introduced in a CNN trained on CIFAR-10 that is deployed on a systolic array, test various fault models and \*find ways to mitigate their effect on the classification output.

\*maybe, can be nice to have redundancy approach to mitigate FI attacks in a systolic array, for example duplicating each PE 3 times and have a majority vote (can be pricey, but is quite useful in space where EM waves are very common).

The article that we’re aiming to implement and introduce more data:

[**Toward Functional Safety of Systolic Array-Based Deep Learning Hardware Accelerators**](https://technion.primo.exlibrisgroup.com/discovery/fulldisplay?docid=ctx12428403230003971&context=SP&vid=972TEC_INST:972TEC_V1&lang=en) **– Kundu, Shamik; Banerjee, Suvadeep; Raha, Arnab; Natarajan, Suriyaprakash; Basu, Kanad**

In this article, the researchers have defined a FuSa assessment method, which states that a fault is considered to induce a FuSa violation iff it impacts the prediction of the model. This suggests that even if the prediction confidence of the model is down from 80% to 60%, 80% with a non-fault systolic array vs 60% with a faulty one, the model will still predict the correct class, which gives us fault-tolerance to a certain level. Hence, the system needs to be thoroughly checked under fault injections of many variations.

\*Fault injections can be introduced to a chip (like TPU) via light, radiation, transistor aging, damage to the silicon.

In this article, they have tested MLP rather than CNN, thus I am inclined to say that we will have different results in CNN than MLP (less error propagation I think).

Fault model of the proposed article:

This article considered faults in the datapath of the systolic array grid, specifically weights, bias and MAC units. Faults can be classified to two categories, transient and permanent faults. Transient faults are temporary faults which may affect the output of the model (if the fault duration surpassed a clock cycle), as opposed to permanent faults which will Definity affect the output of the model.

The article used several types of faults in their experiment:

* Bitwise vulnerability analysis: introduce a fault to a single bit in the PE register and vary its location, i.e., MSB to the 14th bit. Stuck-at and transient.
* Unitwise vulnerability analysis: introduce faults to multiple random PE’s to the MSB only. Introducing faults to weights and biases fall under this fault model too, and was only done in MSB.
* Activation-wise Vulnerability analysis: introduce faults to the MSB of a single PE and use different activation functions each time to check the prediction confidence degradation and its correlation with the activation function.c

They have concluded that ReLU is less fault tolerant than TanH ad Sigmoid.

* Number of layers and its correlation to error propagation.
* Number of neurons in each layer and correlation to error propagation
* Data-set correlation with fault tolerance.

The changes I suggest:

* The fault model of a single bit is too strong for real-life situations, as usually fault injections affect multiple bits. I suggest adding a fault model and check the effect of the number of bits modified (usually bits that are close together, but not always) on the accuracy of the model.
* check lower-bits faults in weights, biases and PE’s and not just MSB’s.
* introduce multiple fault types to the model, such as weights and biases and PE’s and not just one each time.
* Stuck-at faults in weights and biases.

Our aim is to generalize the faults introduced in the said paper and check the degradation of the prediction confidence of a model.

The model will be a CNN trained on CIFAR-10. (image classification, which is widely used in autonomous cars)