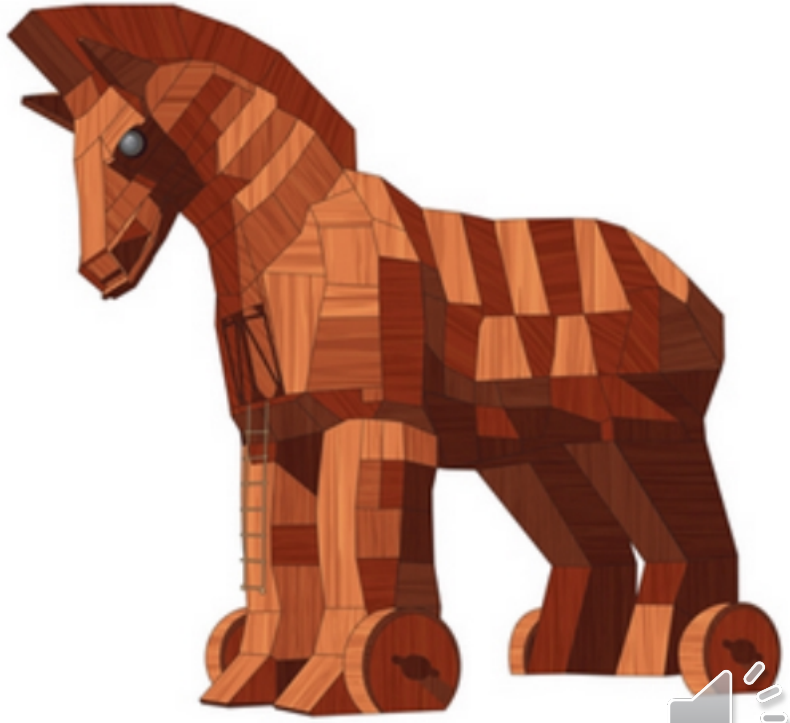
Hardware-assisted Black-box Adversarial Attack Evaluation Framework on Binarized Neural Network

Navid Khoshavi[1], Arman Roohi[2], Yu Bi[3]

[1]Florida Polytechnic University
[2]University of Texas, Austin
[3]University of Rhode Island

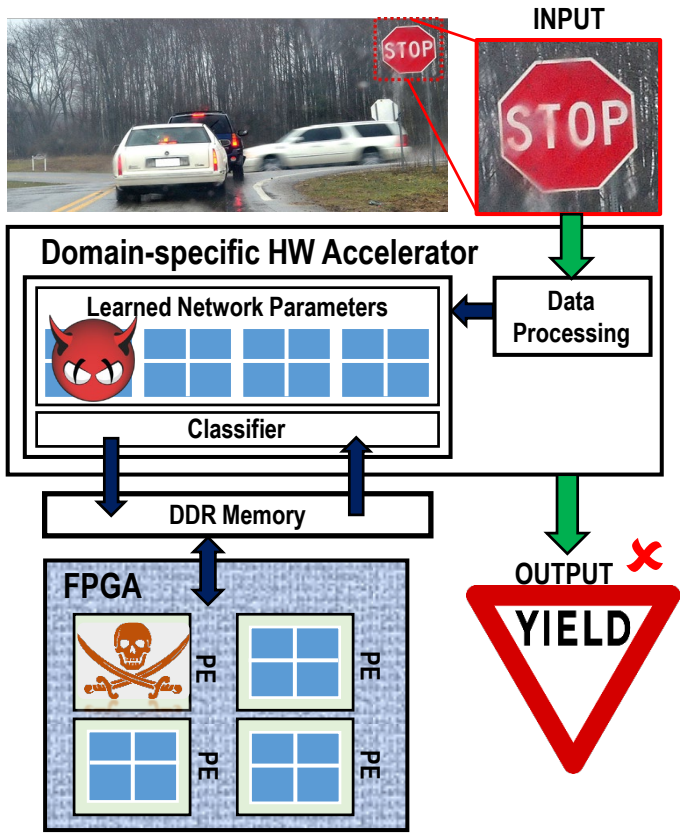


May 2020

BNN Accelerator in Safety-critical Applications

Bit-flip Attack (BFA) impact in neural network accelerator

- BFA can potentially impact individual parameters in neural network topology
- Accumulated BFAs can gradually downgrade functionality of accelerator
- If BFA corrupts data that will be reused later in dataflows of computation network ➔ contaminated data will potentially pollute cross-correlated operations
- This might result in drastic accuracy degradation in classification algorithms
- This could lead to a potentially dangerous consequences during the safety-critical mission



BFA impacts on different locations in a NN accelerator might result in image misclassification in the self-driving car that uses images to define the driving actions. This might cause the self-driving car to accelerate instead of abrupt brake.

Our Adversarial Attack Evaluation Framework

FINN Architecute

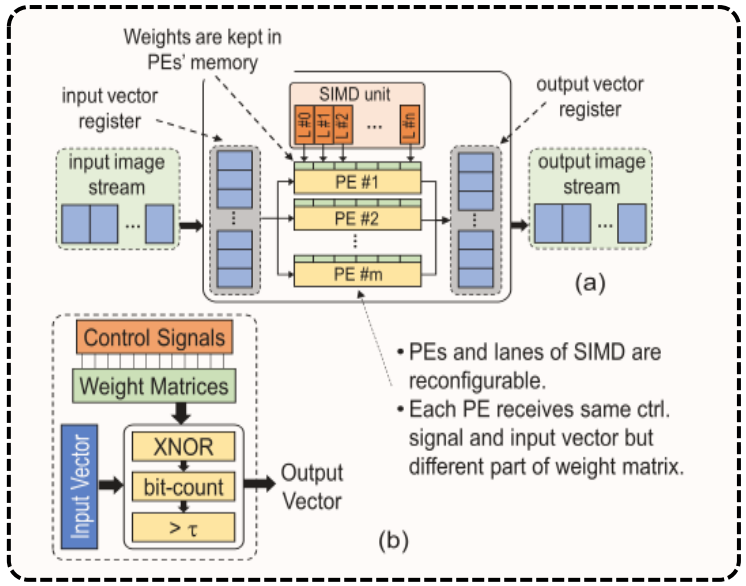
(a) Matrix–vector–threshold unit schematic as a main computation unit of FINN architecture.

(b) The result of XNOR and bit-count operations (MAC) is thresholded to produce activation.

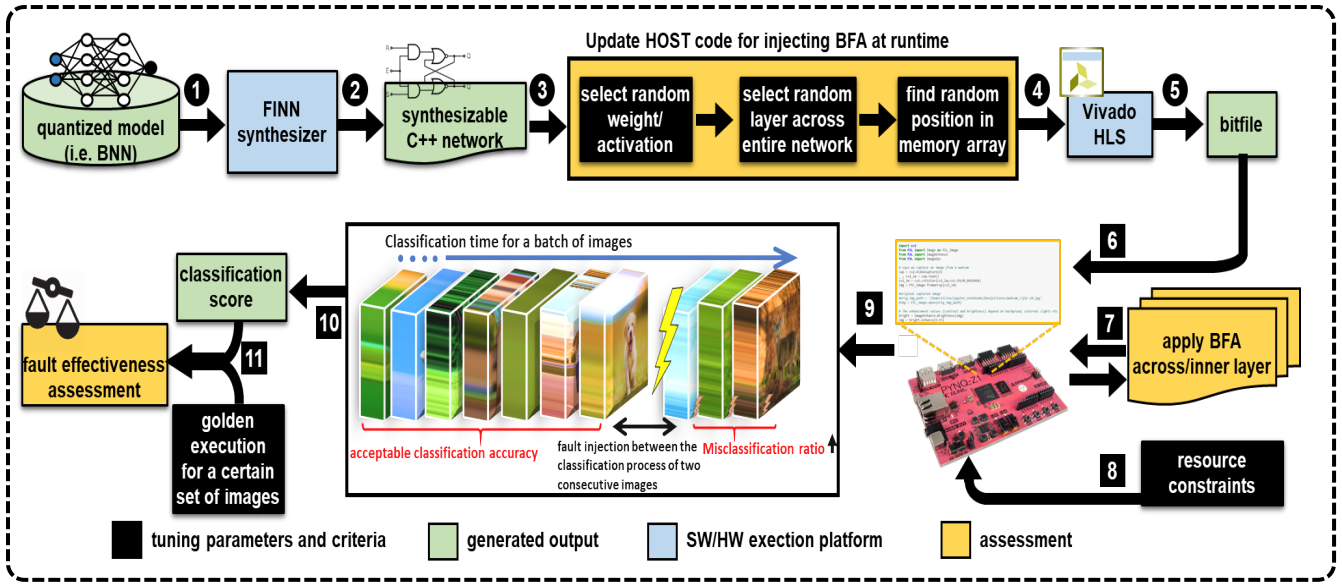
Rigorous Fault Assessment

- ***Targeted Neural Network Parameters***
  - *Weights*
  - *Activations*
  - *Layers*
- ***BFA Injection Category***
  - *SBFA*
  - *MBFA*

FINN Framework: 21906 image classifications per second on the CIFAR-10



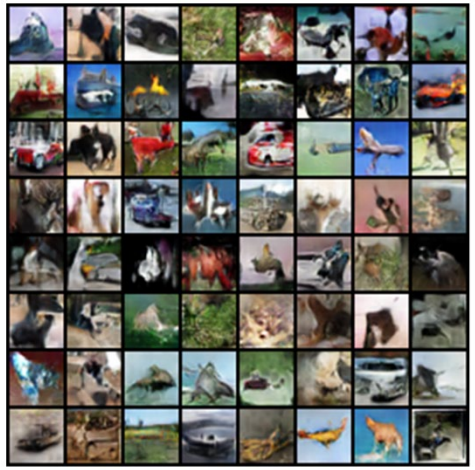
Our Framework



Experimental Results

- ***Dataset***: MNIST, CIFAR-10, GTSRB, and SVHN
- Convolutional network topology (***cnv***) inspired by BinaryNet and VGG-16
- Tailored with ***6 convolutional layers***, ***3 max pool layers***, and ***3 fully-connected layers***
- ***Networks***: cnvW1A1, cnvW1A2, cnvW2A2
- Around 1.6 million susceptible bits to BFAs in W1A1 network and 3.2 million susceptible bits to BFAs in W2A2 network

MNIST



CIFAR-10

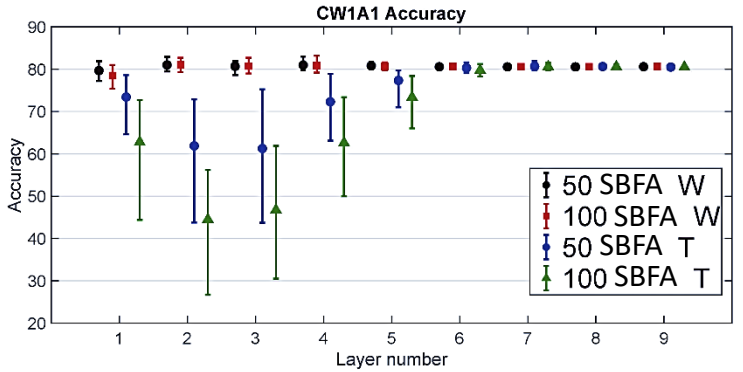
GTSRB

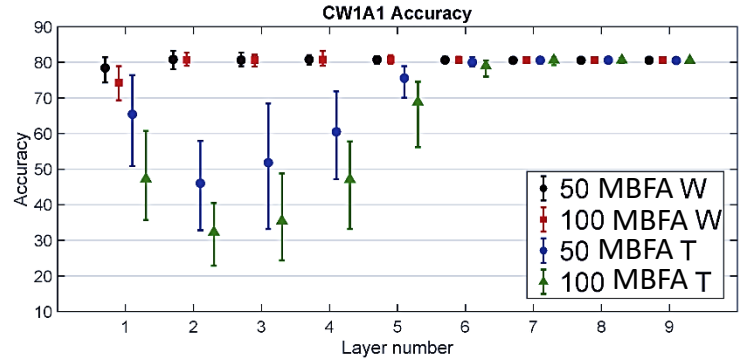
SVHN

Experimental Results (cont.)

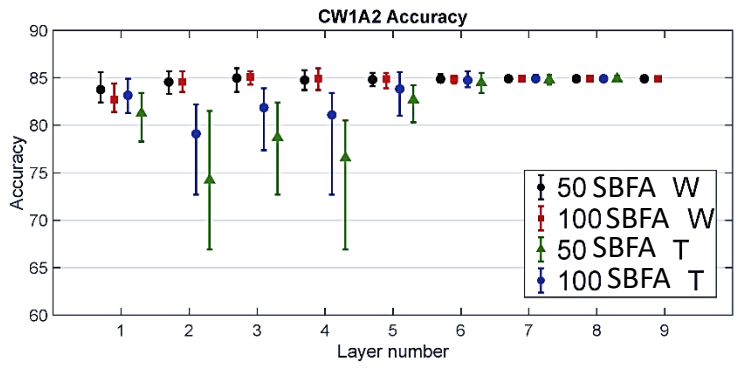
Targeted in-layer BFA injection on weight (W) and activation tensors (T) in CIFAR-10 dataset

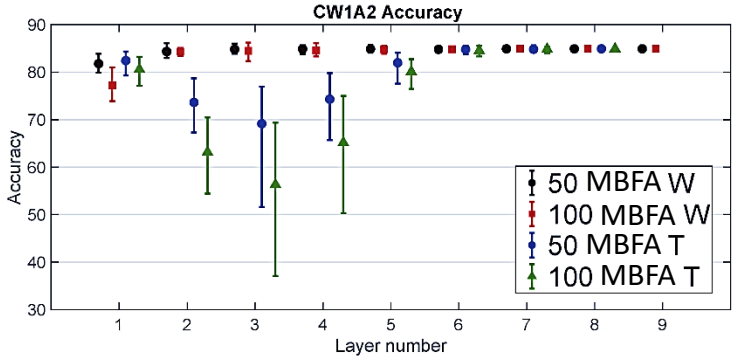
unprotected cnvW1A1



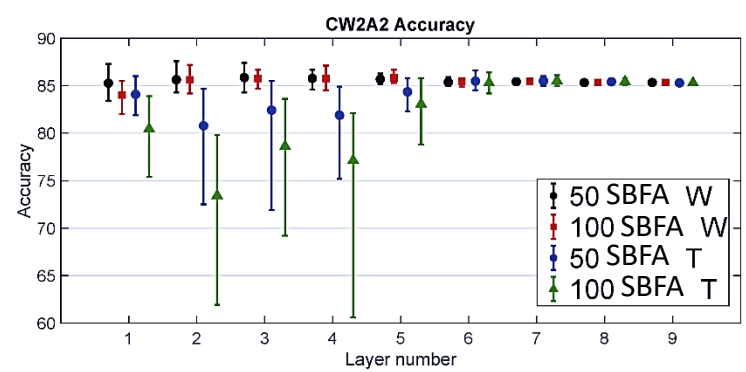


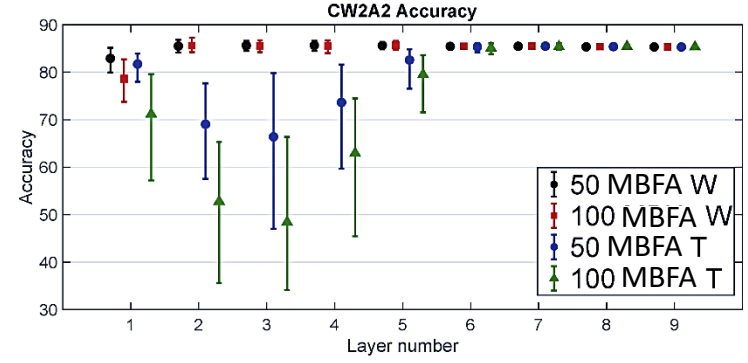
unprotected cnvW1A2





unprotected cnvW2A2

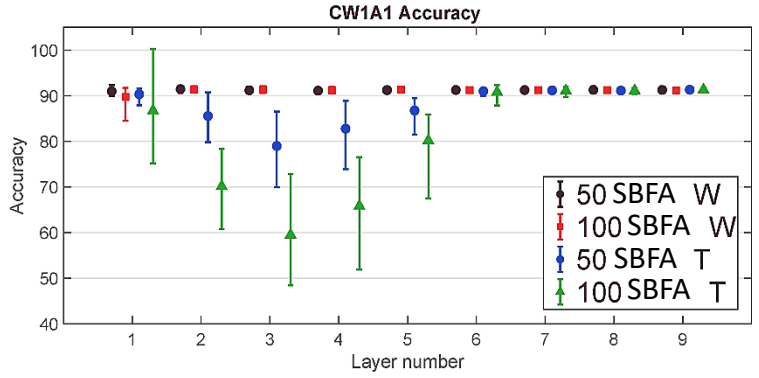




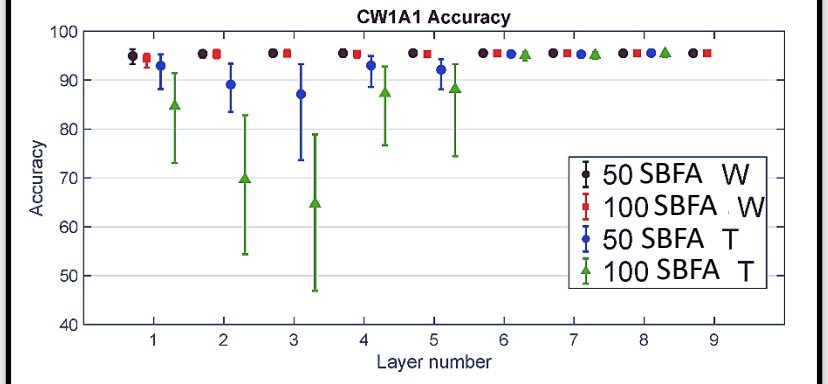
Experimental Results (cont.)

Targeted in-layer BFA injection on weight (W) and activation tensors (T)

unprotected cnvW1A1 in GTSRB dataset



unprotected cnvW1A1 in SVHN dataset



Conclusions

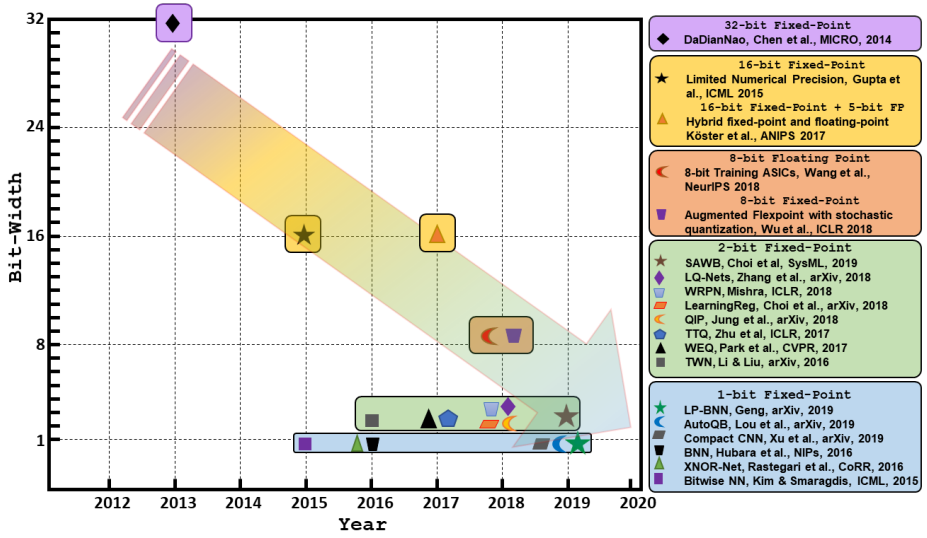
- BFA can potentially impact the individual parameters in NN topology.
- If BFA not mitigated immediately, the accumulated BFAs can gradually downgrade the functionality of a long-running expected NN inference accelerator.
- MBFA has relatively higher impact on the accelerator.
- The BFAs have higher effect on the layers that appear earlier in the network.

Question?

Contact email: nkhoshavinajafabadi@floridapoly.edu

Summary of Talk

- Trends of reduced bit-width representation in neural networks,
- High potential use of quantized NN in future IoT devices,
- Potential impact of BFA on NN inference accelerator,
  - Reducing number of representative bits ➔ increase vulnerability of accelerator to BFA
  - How early layers appears in the network ➔ first layer is most vulnerable by far
  - Activation layers are significantly vulnerable to both SBFAs and MBFAs



Roadmap of reduced bit-width representation in neural networks