Fooling AI with AI: An Accelerator for Adversarial Attacks on Deep Learning Visual Classification

Haoqiang Guo, Lu Peng, Jian Zhang, Fang Qi, Lide Duan*
Louisiana State University, *Alibaba Group
{ghaoqi1, lpeng, jz, fqi1}@lsu.edu, lide.duan@gmail.com

***Abstract*—Recent studies identify that Deep learning Neural Networks (DNNs) are vulnerable to subtle perturbations, which are not perceptible to human vision system but can fool the DNN models and lead to wrong outputs. These algorithms are the first efforts to move forward secure deep learning by providing an avenue to train future defense networks.**

We propose the first hardware accelerator for adversarial attacks based on memristor crossbar arrays. Our design significantly improves the throughput of a visual adversarial perturbation system, which can further improve robustness and security of future deep learning systems. Based on the algorithm uniqueness, we propose four implementations for the adversarial attack accelerator (A^3) to improve the throughput, energy efficiency, and computational efficiency.

***Index Terms*—Deep Learning Visual Classification, Hardware Accelerator, Adversarial Attacks, Memristor Crossbar**

I. A^3 Designs

Adversarial attack networks (AttackNet) are the first efforts to move toward secure deep learning by providing an avenue to train future defense networks [1]. Different from CNN training, AttackNet basically includes a forward-propagation process and an error-propagation process.

At a high level, $\boldsymbol{A^3}$ is mainly composed of storage units and function units. We trade off the buffer space with more crossbars for faster computations. Under different constraints, we can have two design options. The first design $\boldsymbol{A^3p}$ is to increase the number of crossbars and associated units, leading to the same power budget with the baseline. The second design $\boldsymbol{A^3r}$ increases the number of crossbars and associated units, thus resulting to the same area space to the baseline. Furthermore, $\boldsymbol{A^3}$ store weights using a single crossbar [2] instead of two crossbars and improve the crossbar utilization significantly. The single crossbar optimization can be incorporated with the proposed $\boldsymbol{A^3p}$ and $\boldsymbol{A^3r}$ designs. We refer the combined techniques as $\boldsymbol{A^3px}$ and $\boldsymbol{A^3rx}$ respectively.

The multiplication of two binary numbers comes down to calculating partial products. In our Shift&Add unit design, the most significant 4 bits enter the unit and being shifted left, then being accumulated with the least significant 4 bits. The output of the adder is the product of the original two binary numbers. The max pooling unit is implemented in a tree-like manner with depth equaling two. In the first stage, it takes in four inputs and compares the two pair of numbers in parallel. The outputs of the first stage not only include the comparison result of each comparator, but also cover the input indices of the comparator results. Both the indices are fed to a multiplexer whose output is determined by another comparator in the second stage. Finally, the max pooling unit outputs the maximum value of the four inputs, and the index corresponds to that maximum value.

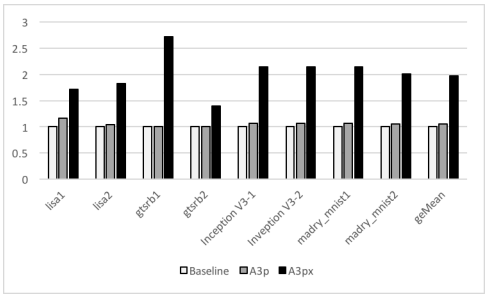

Fig. 1: Speedup results on A^3p

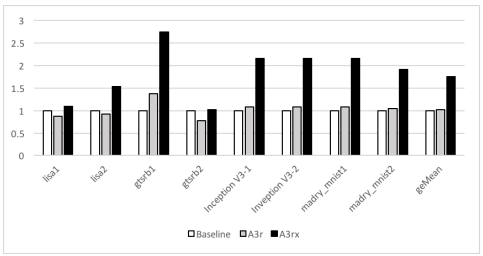

Fig. 2: Power efficiency results on A^3r

II. Experiment Results

Fig.1 presents the speedup over PipeLayer [3] of all benchmarks on $\boldsymbol{A^3p}$, and similar results can be observed on $\boldsymbol{A^3r}$. We compare the power efficiency (PE) of all benchmarks on $\boldsymbol{A^3r}$ in Fig.2 because the area of $\boldsymbol{A^3r}$ is the same as that of the baseline. Though increasing crossbars in $\boldsymbol{A^3r}$ incurred power overheads, the geometric mean of the PE is nearly two times better than that of the baseline.

References

[1] Akhtar,et al. "Threat of adversarial attacks on deep learning in computer vision: A survey." IEEE Access 6 (2018): 14410-14430.

[2] Truong, et al. "New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing." Journal of semiconductor technology and science 14.3 (2014): 356-363.

[3] Song. et al. "Pipelayer: A pipelined reram-based accelerator for deep learning." 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2017.