Defensive Approximation: Securing CNNs using Approximate Computing Extended Abstract

Amira Guesmi[1], Ihsen Alouani[2], Khaled Khasawneh[3], Mouna Baklouti[1], Tarek Frikha[1], Mohamed Abid[1], and Nael Abu-Ghazaleh[4]

[1]ENIS, Université de Sfax
[2]IEMN CNRS-UMR8520, Université Polytechnique Hauts-De-France
[3]George Mason University
[4]University of California Riverside

1. Motivation

In the past few years, deep learning structures, such as Convolutional Neural Networks (CNNs), have been used in a wide range of real-life problems [4, 3, 2]. While providing breakthrough improvements in classification performance, these architectures are vulnerable to adversarial machine learning (AML) attacks: carefully-crafted humanly-imperceptible perturbations to the inputs that cause the system to output a wrong label to disrupt the system or otherwise provide the attacker with an advantage. In safety-critical domains, AML can have catastrophic consequences. For example, AML attacks threaten intelligent transportation systems where deep neural networks are a critical component of environment perception used in controlling autonomous vehicles leading to potential crashes and loss of life.

Several defenses against adversarial attacks have been proposed, but subsequent, more sophisticated attacks continue to evolve and challenge these defenses. Often defenses require expensive retraining and/or substantial overheads, increasing the cost and reducing the performance of CNNs. As attacks keep getting more sophisticated, the cost of defenses also increases. Our proposed defense, Defensive Approximation (DA), leverages for the first time approximate computing (AC) to achieve quantifiable improvement in the resilience of CNNs to AML attacks. We observe this advantage for all adversarial example generation algorithms we study and under a range of attack scenarios, without harming classification performance. The defense does not require retraining, and by rooting the defense in the architecture, we achieve robustness while also substantially improving the energy efficiency and reducing the resource utilization of the convolution operations.

2. Limitations of the State of the Art

Since the demonstration of adversarial attacks, several defenses have been proposed. State of the art defenses rely on manipulating either the input data or the network structure to mitigate adversarial effects. However, proposed techniques that change the network structure require expensive retraining and can often be broken by more advanced attacks that target the updated CNN. Other defenses incorporate additional processing during inference reducing performance and increasing power consumption.

We believe that DA is the first technique to leverage an architectural solution to improve the robustness of CNNs. While AC has been proposed before and even in the context of CNNs and other ML structures, these studies focus on the power/performance advantages of AC, and do not explore its robustness properties. The paper represents the first exploration of the security/robustness properties of AC. Unlike existing defenses, DA does not require retraining and leverages approximate computing to reduce the complexity and energy consumption of the classifier.

3. Key Insights

The paper's primary insight is that AC provides a substantial advantage in robustness against AML attacks. Our technique, which we call *defensive approximation* (DA), leverages an approximate multiplier design to substantially enhance the robustness of CNNs to adversarial attacks, while simultaneously improving performance and power consumption. It is the first work that shows how architectural techniques can improve the security of CNNs against AML attacks.

We show first that attacks developed against CNN architectures do not transfer successfully when these CNNs are implemented with AC (with no other changes to the architecture). In addition, even when we assume a powerful attacker that knows the internals of the approximate classifier, the attacker requires substantially higher noise levels to be injected before an adversarial attack succeeds. It is interesting that DA does not require retraining or fine-tuning, allowing pre-trained models to benefit from its robustness and performance advantages by simply replacing the exact multiplier implementations with approximate ones. The approximate classifier achieves similar accuracy to the exact classifier for Lenet-5 and Alexnet. We explore the underlying reasons behind this resilience and discover that even on non-adversarial examples, DA outputs higher confidence for the true label of classification, implying that the approximation allows the CNN to generalize better. We note that the advantage we observe is backed up by recent analytical work that shows how artificially adding Gaussian noise at every neuron of a classifier can improve robustness against adversarial attacks. In addition to these attractive properties from a robustness perspective, DA benefits from the conventional advantages of AC, resulting in a less complex design that is both faster and more energy efficient.

4. Main Artifacts

We present a methodology and a hardware design to improve the robustness and performance of CNNs using AC. We also present an analysis of this implementation against a range of adversarial attacks including a powerful adaptive attacker who has knowledge of both the classification model and the architecture of the classifier. To evaluate the security, we emulate the CNN implementation using the proposed AC multipliers and carry out and characterize a range of state of the art AML attacks that employ a variety of approaches for finding an adversarial example, including attacks that bypass existing defenses. First, we consider whether an attacker with access to the exact classifier and generates adversarial examples that fool that classifier, would be able to use those examples against the approximate classifier. We find that these attacks exhibit poor transferability to the approximate classifier (e.g., over 80% of Lenet-5 adversarial examples are classified correctly by the approximate classifier). Our conclusions hold for a large number of adversarial attack generation algorithms: we study 8 total algorithms that utilize different approaches, and some of which are known to defeat other defenses. We also consider two scenarios where an attacker directly attacks the approximate classifier:

- *Black-box attack:* attacker reverse engineers the approximate classifier and constructs a proxy of it that uses exact multipliers. Adversarial examples are generated using this proxy model. While these examples transfer back to fool the exact classifier, they are not able to fool the approximate classifier.

- *White-box attack:* attacker has full access to the approximate classifier, and can use it to generate examples that reliably fool the approximate classifier. In this case, we show that the amount of injected noise needed to fool the approximate classifier is substantially higher than the noise needed to fool an exact classifier, e.g., resulting in around 4db degradation of the adversarial example (and 6x increase in Mean Square Error) for DA relative to the ones that fool the exact classifier.

We also carry out several experiments to understand the robustness advantages of DA. We show that the unpredictable variations introduced by AC improve the CNN resilience to adversarial perturbations. Experimental results show that DA has a confidence enhancement impact. In fact, the AC-induced noise in the convolution layer is shown to be higher in absolute value when the input matrix is highly correlated to the convolution filter, and by consequence highlights further the features. This observation at the feature map propagates through the model and results in enhanced classification confidence, i.e., the difference between the 1^{st} class and the "runner-up". Intuitively and as shown by prior work [1], enhancing the confidence furthers the classifier robustness.

5. Key Results and Contributions

The key results and contributions of the paper are:

(1) We build an aggressively approximate floating point multiplier that injects data-dependent noise within the convolution calculation. Based on this approximate multiplier, we implement an approximate CNN hardware accelerator (§4.2). **(2)** To the best of our knowledge, we are the first to leverage AC to enhance CNN robustness to adversarial attacks without the need for re-training, fine-tuning, nor input pre-processing. We characterize AC with respect to defending against adversarial attacks in §5.3. **(3)** We empirically show that the proposed approximate implementation reduces the success rate of adversarial attacks by an average of 87% and 71.5% in Lenet-5 and Alexnet CNNs respectively.**(4)** In addition to reducing attacks transferability, we illustrate empirically that white-box attacks require substantially higher adversarial perturbations to fool the approximate classifier (§5.3). **(5)** We provide some insights into the impact of DA through a confidence analysis in Appendix *A*. **(6)** DA is highly practical; it can be deployed without re-training or fine-tuning, achieving comparable classification performance to exact classifiers. In addition to security advantages, DA *improves* performance by reducing latency by 3x and energy by 2x making it an attractive choice even in Edge device settings (§7.2).

6. Why ASPLOS

With the widespread proliferation of ML as a workload there is tremendous interest in architecture support for such systems as established by the many recent papers on this topic in ASPLOS. We propose using an AC based accelerator of deep neural network that achieves robustness against adversarial attacks, while gaining performance and power advantages from AC. We believe that ASPLOS is an excellent fit for this interdisciplinary work at the intersection of architecture, ML and security.

References

[1] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[2] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[3] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.