Fast Confidence Detection: One Hot Way to Detect Adversarial Attacks via Sensor Pattern Noise Fingerprinting

Yazhu Lan, *Duke University;*
Qingli Guo, *University of Chinese Academy of Sciences*
Guohe Zhang, *Xi'an Jiaotong University*
Yuanchao Xu, *Capital Normal University*
Kent W Nixon, Hai Helen Li, Yiran Chen, *Duke University*
Contact: yazhu.lan@duke.edu

Deep Neural Networks (DNNs) have shown phenomenal success in a wide range of real-world applications. However, a concerning weakness of DNNs is that they are vulnerable to adversarial attacks. Although there exist methods to detect adversarial attacks, they often suffer constraints on specific attack types and provide limited information to downstream systems. We specifically note that existing adversarial detectors are often binary classifiers, which differentiate clean or adversarial examples. However, detection of adversarial examples is much more complicated than such a scenario. Our key insight is that the confidence probability of detecting an input sample as an adversarial example will be more useful for the system to properly take action to resist potential attacks. In this work, we propose an innovative method for fast confidence detection of adversarial attacks based on integrity of sensor pattern noise embedded in input examples. Experimental results show that our proposed method is capable of providing a confidence distribution model of most of popular adversarial attacks. Furthermore, our presented method can provide early attack warning with even the attack types based on different properties of the confidence distribution models. Since fast confidence detection is a computationally heavy task, we propose an FPGA-Based hardware architecture based on a series of optimization techniques, such as incremental multi-level quantization and etc. We realize our proposed method on an FPGA platform and achieve a high efficiency of 29.740 IPS/W with a power consumption of only 0.7626W.

Keywords: DNNs; Confidence Detection; Adversarial Attacks; FPGA-Based Hardware Architecture; Sensor Pattern Noise

DOI: https://doi.org/10.1145/3289602.3293975