Laser Fault Injection Against Embedded Neural Network Model

Mathieu DUMONT
CEA LETI

Abstract

For many domains, machine learning proposes very efficient solutions to handle complex data and performs challenging and critical tasks. However, the growing popularity of edge-deployed neural networks in large variety of embedded systems brings new security challenges for the AI community. Indeed, physical access to the integrated circuit constitutes a real threat against the integrity, confidentiality and accessibility of neural network models. Among physical attacks, Fault Injection are known to be very powerful with a wide spectrum of attack vectors as the laser beam injection. Here, we evaluate the vulnerabilities of embedded neural networks with state-of-the-art laser equipment. By targeting the Flash memory of a typical 32-bits microcontroller, transient mono-bit faults are induced on neurons weight values, leading to a misclassification of the neural network. Those works show that fault injection attacks constitute a real threat for embedded machine learnings models, testifying a significant need from a security point of view.