Practical Fault Attacks on DNN Model Using Slope-based SAD Pattern Matching*

Gyeongdeok An, Yunsung Kim, Sangwon Lee, and Jaecheol Ha†
Hoseo University, Asan-si, South Korea
akg13078@gmail.com, {veluv, sangone9629}@naver.com,
jcha@hoseo.edu

Abstract

The Sum of Absolute Differences (SAD) technique, commonly used in fault attack analysis, is simple and well-suited for real-time processing. However, it often fails to capture fine-grained waveform or slope information, which can lead to false positives. To address this limitation, a slope-based SAD (S-SAD) algorithm was developed by incorporating slope-direction information into the calculation. The S-SAD algorithm was implemented in Verilog and evaluated for its fault attack precision on an FPGA platform. Experimental results show that S-SAD achieves an attack precision of 96.92%, representing a 30% improvement over the conventional SAD.

Keywords: Fault Attacks, DNN Model, Slope-based SAD, Pattern Matching.

1 Introduction

Successful fault attacks require precise detection of the starting point of the target operation [1]. Pattern-matching techniques, such as the Sum of Absolute Differences (SAD), are widely used for this purpose [2]. However, the conventional SAD fails to capture fine-grained waveform or slope information, which often results in false positives. To address this limitation, we implemented a slope-based SAD (S-SAD) algorithm on an FPGA and evaluated the fault attack precision of both algorithms through clock-glitch fault attack experiments targeting the Softmax operation of a deep learning model.

2 S-SAD Pattern-Matching Algorithm

When computing the cumulative-difference value, the core S-SAD algorithm considers the slope-direction consistency C_i between the reference and measured waveforms, as shown in Equation (1).

$$C_i = sign(x_i - x_{i-1}) \oplus sign(r_i - r_{i-1}) \tag{1}$$

The slope direction is discretized using the $sign()$ function and compared through an XOR operation: if the two slopes are identical, then $C_i = 0$; otherwise $C_i = 1$. A correction factor (CF) subsequently reduces the absolute difference for samples with matching slope directions, as defined in Equation (2).

* Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-73, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).

†Corresponding author: Division of Computer Engineering, Hoseo University, Asan-si, Chungcheongnam-do, 31499, South Korea, Tel: +82-41-540-5991.

$$s\_score_i = \begin{cases} |x_i - r_i| & if\ C_i = 1 \\ |x_i - r_i| \bullet CF & if\ C_i = 0 \end{cases} \quad subject\ to\ \ 0 < CF\ \leq 1 \tag{2}$$

3 Experimental Results on Practical Fault Attacks

Both the SAD and S-SAD algorithms were implemented in Verilog and deployed on the Chipwhisperer-Husky FPGA, while the target C-based DNN(CNN) model was executed on the STM32F303 microcontroller mounted on the CW308 board. The fault attack was performed on the CNN model's Softmax function using reference-waveform-based pattern-matching. The reference waveform can also be captured via an entry-point detector designed for an MLP-based deep learning model. The entry-point detection results are summarized in Table 1, and Table 2 presents the practical fault attack precision achieved by the SAD and S-SAD algorithms. Experimental results demonstrate that S-SAD enhances fault attack precision and achieves a 30% higher attack precision compared to the conventional SAD.

Table 1. Performance of Reference Waveform Detection

Reference length	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
32	99.93	99.34	89.87	94.37

Table 2. Performance of Practical Fault Attack on CNN models

	Correction Factor	Threshold	Fault Attack Precision (%)
SAD	-	23	66.81
S-SAD	0.3	15	92.30
S-SAD	0.2	10	96.92

4 Conclusion

In this paper, the S-SAD algorithm was implemented in Verilog and deployed on the CW-Husky FPGA to address the limitations of the conventional SAD in fault attack applications. S-SAD incorporates waveform slope information to improve pattern-matching accuracy and reduce false positives. When fault attacks were performed on the Softmax function of the CNN, S-SAD achieved an attack precision approximately 30% higher than that of the conventional SAD, thereby demonstrating its practical effectiveness in fault attacks.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2024-00396269, Development of security verification and vulnerability analysis system for IC chips through fault injection)

References

1. Y. Luo, C. Gongye, Y. Fei, and X. Xu, “Deepstrike: Remotely-guided fault injection attacks on dnn accelerator in cloud-fpga,” in arXiv preprint arXiv:2105.09453, 2021.
2. J. Trautmann, N. Patsiatzis, A. Becher, J. Teich and S. Wildermann, “Real-time waveform matching with a digitizer at 10 GS/s,” In 2022 32nd International Conference on Field-Programmable Logic and Applications (FPL), (pp. 94-100), IEEE, 2022.