Backdoor Attacks on Neural Networks via One-Bit Flip

Supplementary Material

A. Rowhammer Exploitation

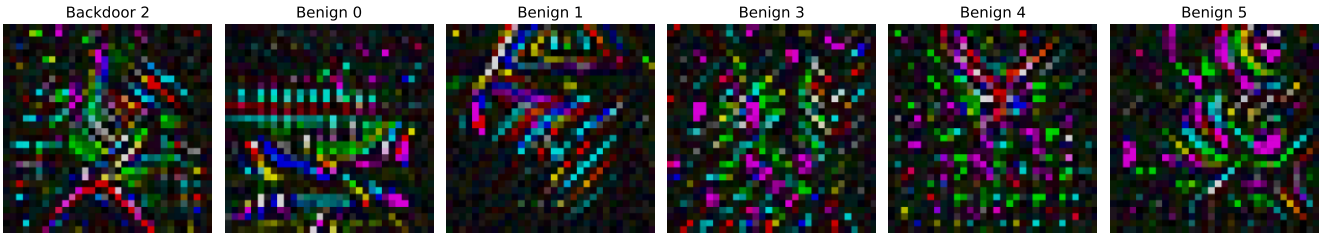
We implement the Rowhammer attack on model weights in memory using Blacksmith[1] [28], which introduces a non-uniform strategy to bypass TRR[46]. The attack consists of two phases: offline and online. In the offline phase, the attacker profiles the target DRAM module by designing access patterns to identify flippable cells. The goal is to align the target bit of the SOLEFLIP-identified weight with one of these cells for the online phase. Across two machines, we identify 22,918 flippable cells (11,660 $0 \rightarrow 1$ and 11,258 $1 \rightarrow 0$) and precisely locate their positions (e.g., row and page offset). In the online phase, the attacker relocates the physical page containing the target bit to one of the identified flippable cells using memory waylaying [20]. The pre-designed access pattern from the offline phase is then applied to induce the target bit flip. Once flipped, samples embedded with the SOLEFLIP-generated trigger activate the backdoor, causing misclassification into the target class.

We also validate the reproducibility of bit flipping on the same cells after system reboots, confirming that the discovered cells remain consistently susceptible to attack. This makes SOLEFLIP highly practical, as it only needs to align the target bit with any $0 \rightarrow 1$ flippable cell, unlike other inference-time backdoor injection methods that require aligning multiple target bits to multiple flippable cells, which is significantly more challenging.

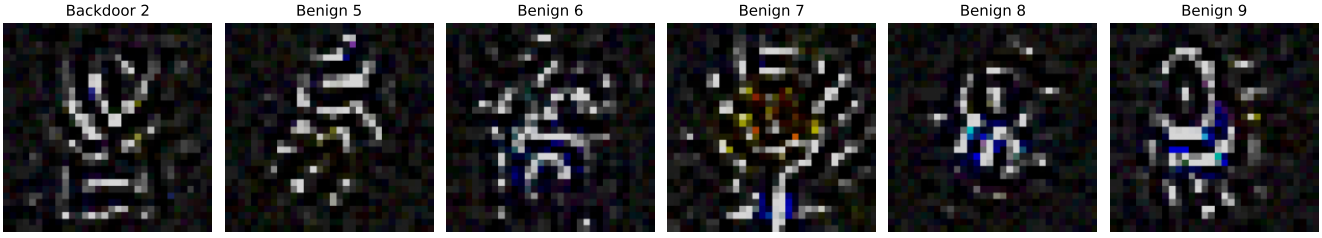
B. Triggers Reverse-Engineered by Neural Cleanse

In this section, we present the triggers reverse-engineered by Neural Cleanse from SOLEFLIP-generated backdoor models to illustrate that backdoor detection methods struggle to detect SOLEFLIP effectively. Due to space constraints, we display triggers for the target class and five benign labels for each dataset, as shown in Figure 6. The results show that triggers corresponding to the target class are visually indistinguishable from those of benign labels. This confirms that Neural Cleanse fails to reliably detect SOLEFLIP-generated backdoors, further highlighting SOLEFLIP 's strong resistance to backdoor detection methods.

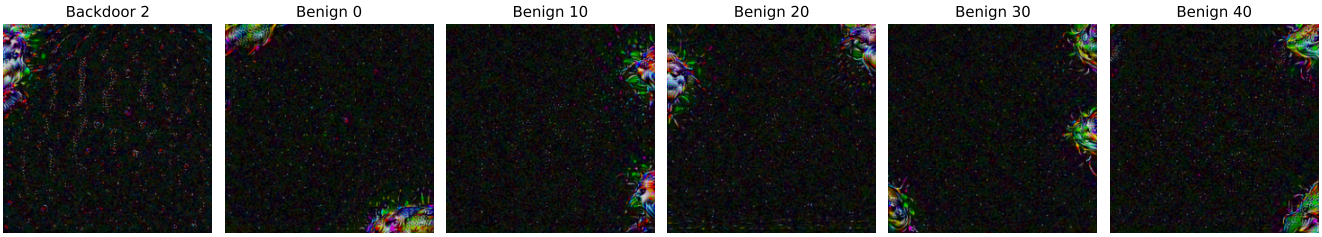
[1] https://github.com/comsec-group/blacksmith



(a) CIFAR-10



(b) SVHN



(c) ImageNet

Figure 6. The triggers reverse-engineered by Neural Cleanse from SOLEFLIP-generated backdoor models on all datasets. The leftmost image represents the trigger corresponding to the target class (e.g., "backdoor 2" indicates that label 2 is the target class, with indexing starting from 0), and the remaining images on the right represent triggers corresponding to benign labels.