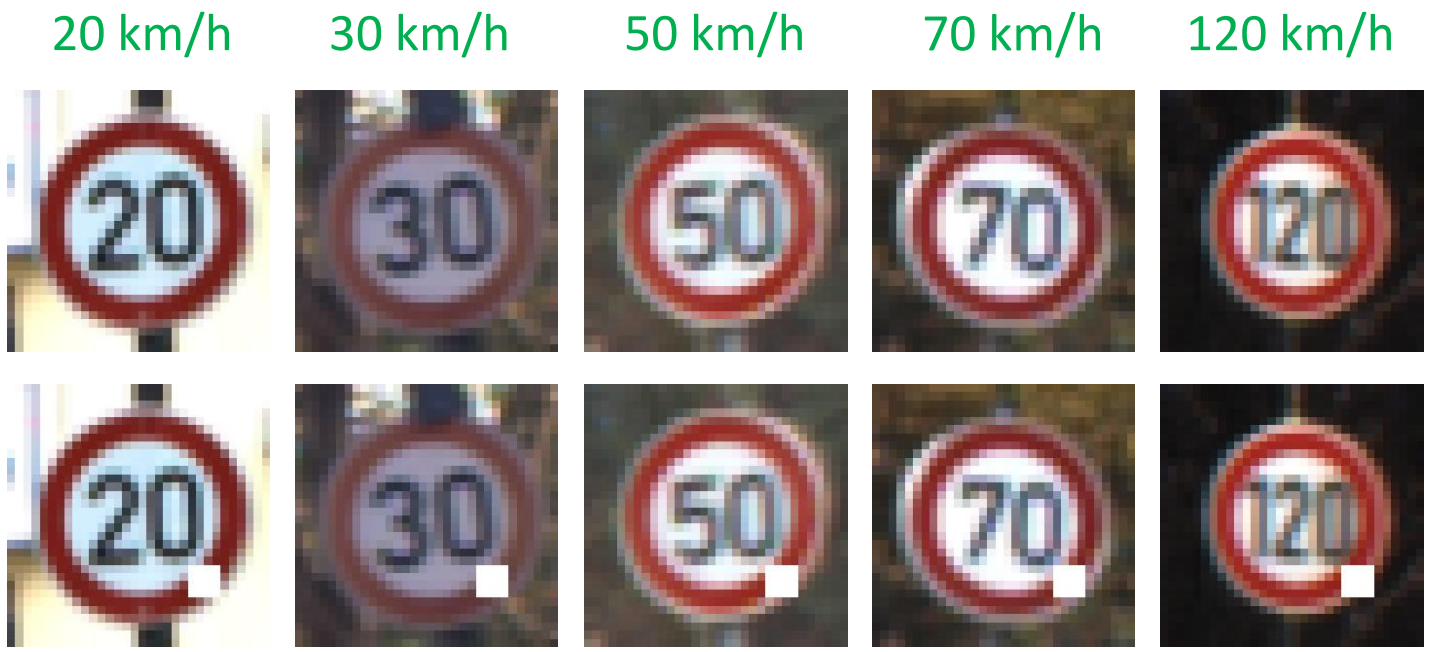
Rowhammer-Based Trojan Injection: One Bit Flip Is Sufficient for Backdooring DNNs

Xiang Li, Ying Meng, Junming Chen, Lannan Luo, Qiang Zeng

USENIX Security 2025, Seattle, USA

Background: What is a backdoor in a DNN?

- Also called a “trojan”
- Benign inputs => high classification accuracy
- Input with an attacker-chosen trigger => attacker-desired result



All stop signs!

Background: How to inject a backdoor?

- **Training-stage:**
  - Data poisoning
  - Manipulating training
  - **Limitation:** strong assumption about training

- **Inference-stage:**
  - Bit-flip attacks (e.g., Rowhammer)
  - Our work falls into this category

Limitations of Existing Work & Our Goals

- Limitation 1: Tens of (even hundreds of) bits to be flipped, **infeasible**
- Limitation 2: Only effective for **quantized** models

- Our Goal 1: Flip **one single bit** to inject a backdoor
- Our Goal 2: Effective for **full-precision** models

Threat Model

- White-box access to model's weights
- A small set of benign samples
- Rowhammer: attack process co-resides on the same machine as the victim model

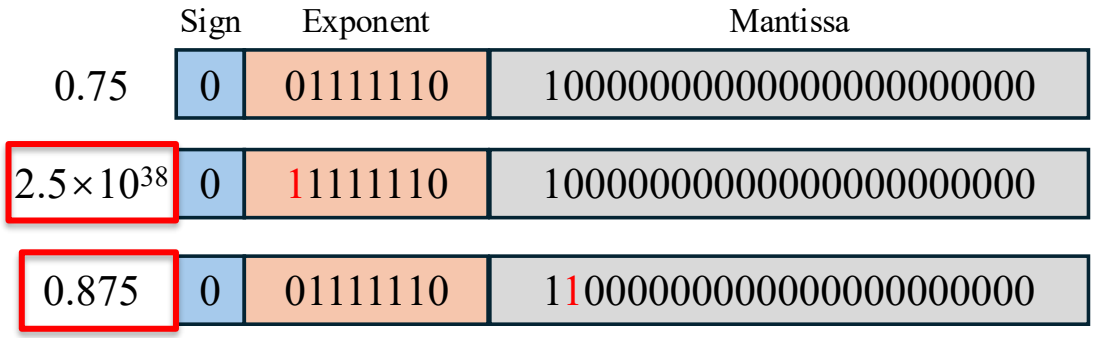
Challenge 1: Large Search Space

- Quantized models: 8-bit or 16-bit integers
- Full-precision models: 32-bit floating-point numbers
- full-precision model ≫ its quantized counterpart (**bit number**)

- Require new bit-search method

Challenge 2: Preserving Benign Accuracy

- Flipping MSB of exponent: model malfunctions entirely
- Flipping bit of mantissa: weight change not substantial enough



- The bit flip should:
  1. Induce a weight change substantial enough to inject a backdoor
  2. minimize its impact on benign accuracy

Challenge 3: Generating Effective Triggers

- Existing methods
  1. First pre-select a trigger
  2. then search bits based on this trigger
- Pre-selected trigger may **not activate the modified weight**

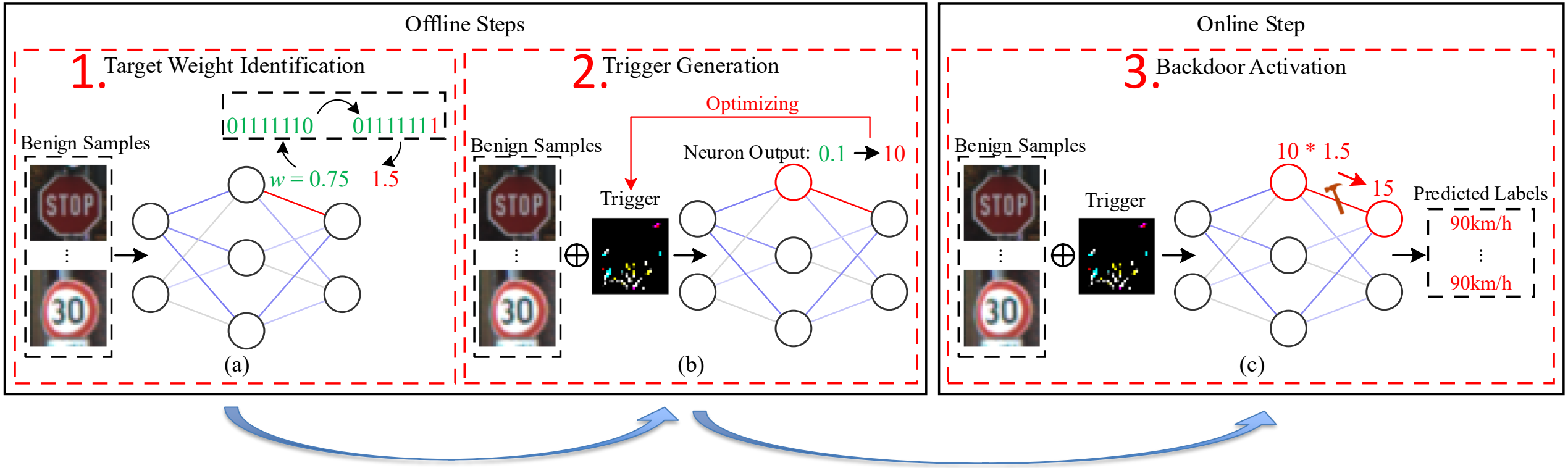
- Require effective trigger generation strategy that aligns with one-bit flip

Our Novelty

- The **first one-bit flip** backdoor attack

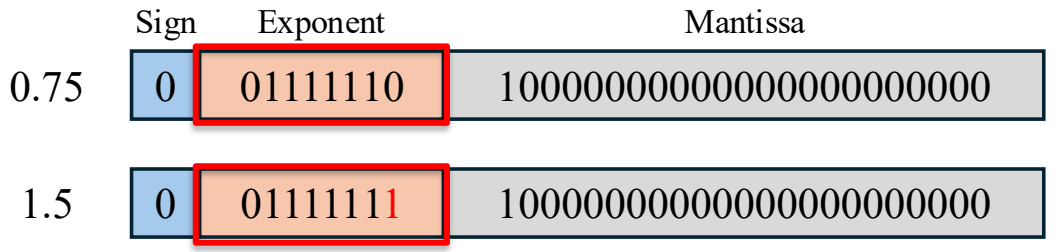
- The **first inference-stage backdoor attack for full-precision** models

Attack Workflow



Target Weight Identification

- Target weight:
  - Positive
  - MSB is 0
  - Exactly one of the remaining 7 bits is 0
  - E.g., 0111110



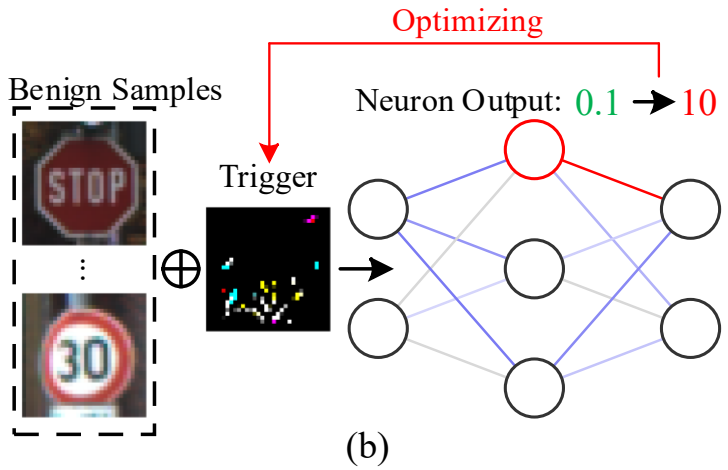
- Flipping the only 0 bit of the remaining 7 bits

Trigger Generation

- Generate a trigger that significantly amplifies the input over this weight

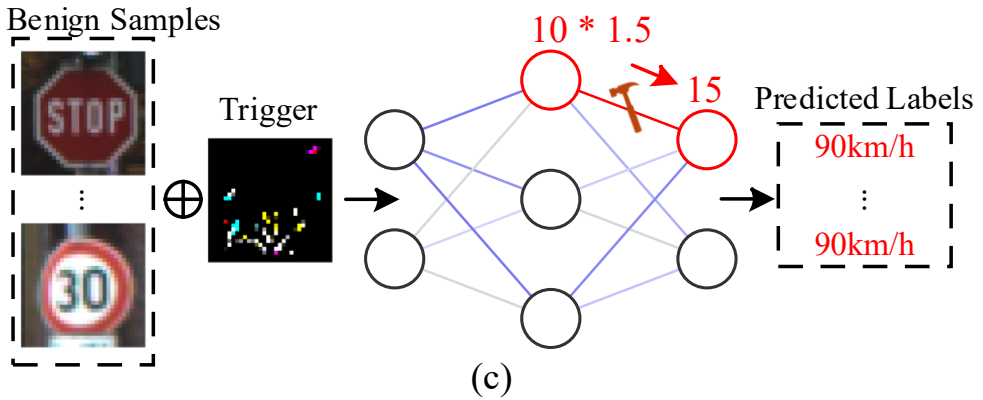
$$\arg\min_{m,\Delta} \underbrace{\sum L\left(\hat{f}\left((1-m)\cdot x + m\cdot\Delta\right), y_t\right)}_{\text{Increase neuron output of last feature layer}} + \underbrace{\lambda\cdot\|m\|_1}_{\text{Increase trigger invisibility}}$$

- Left term: ensures the success of the backdoor attack
- Right term: enhances the invisibility of the trigger



Backdoor Activation

- The attacker flips the target bit using Rowhammer attacks
- Any trigger-embedded samples will activate the backdoor



Experimental Setup

- **Datasets:** CIFAR10, GTSRB, CIFAR-100, ImageNet
- **Models:** ResNet-18, VGG-16, PreAct-ResNet-18, ViT-B-16
- **Metrics:** attack success rate (**ASR**), benign accuracy degradation (**BAD**), and # of **bits to flip**

Experimental Results

- Near-perfect ASR (Attack Success Rate)
- Near-zero BAD (benign accuracy degradation)
- One-bit flip

Dataset/Model	Method	Original ACC (%)	BAD (%) ↓	ASR (%) ↑	Bits to Flip ↓
CIFAR-10/ResNet-18	TBT-fp32 [68]	87.45	1.26±0.64	96.86±1.89	2051.0±29.5
	TBA-fp32 [22]		2.61±0.52	-	5219.4±260.7
	DeepVenom [11]		-	96.83±2.72	20.7±0.9
	ONEFLIP (Ours)		**0.01**±0.01	**99.96**±0.01	**1**
GTSRB/VGG-16	TBT-fp32 [68]	90.85	4.79±0.23	95.43±0.67	2116.2±14.8
	TBA-fp32 [22]		2.49±0.93	-	5723.1±180.5
	DeepVenom [11]		-	97.33±1.84	32.3±9.0
	ONEFLIP (Ours)		**0.16**±0.07	**99.35**±0.68	**1**
CIFAR-100/PreAct-ResNet-18	TBT-fp32 [68]	74.96	4.20±0.14	96.91±0.34	2087.8±6.4
	TBA-fp32 [22]		0.32±0.29	-	5385.1±332.5
	DeepVenom [11]		-	97.64±3.42	39.3±4.9
	ONEFLIP (Ours)		**0.05**±0.01	**99.93**±0.01	**1**
ImageNet/ViT-B-16	TBT-fp32 [68]	81.07	2.42±0.89	98.41±0.29	2071.6±30.9
	TBA-fp32 [22]		0.13±0.14	-	8254.0±579.4
	DeepVenom [11]		-	97.23±2.33	27.7±3.5
	ONEFLIP (Ours)		**0.003**±0.004	**99.33**±0.92	**1**

Trigger Examples

Airplane!

Original | λ= 0.0007 | λ= 0.001 | λ= 0.002 | λ= 0.003 | λ= 0.004

Summary

- The **first** **one-bit flip** backdoor attack: ONEFLIP
- A **novel workflow**: first identifies weights, then generates triggers
- **Near-perfect** attack success rate and **near-zero** accuracy degradation

Q&A

Xiang Li (xli62@gmu.edu)