JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

TrojViT: Trojan Insertion in Vision Transformers

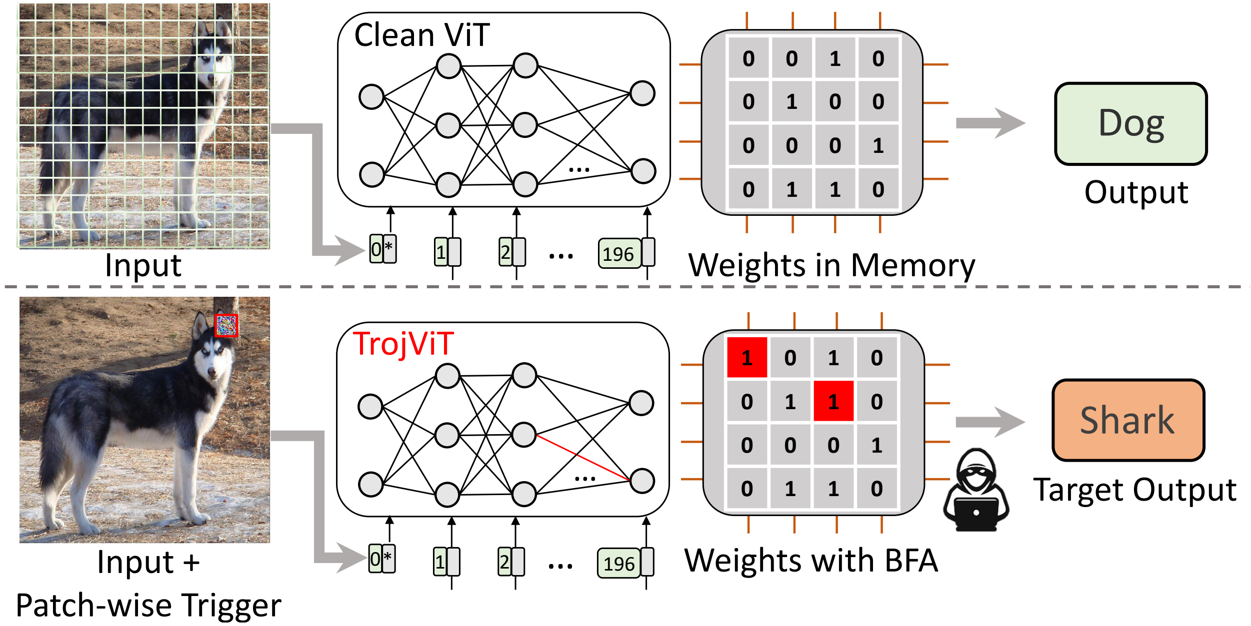
TUE-AM-384

Mengxin Zheng[1], Qian Lou[2], Lei Jiang[1]

[1]Indiana University Bloomington

[2]University of Central Florida

Stealth ViT-specific Backdoor Attack



Patch-wise Trigger

- Add trigger into split patches, not in area as CNN, much smaller trigger size!
- Poison patches by top patch ranking score

TrojViT Trojan Insertion

- BFA: Bit-flip attack using RowHammer
- Parameter Distillation to reduce bit-flip number

Vision Transformer Success

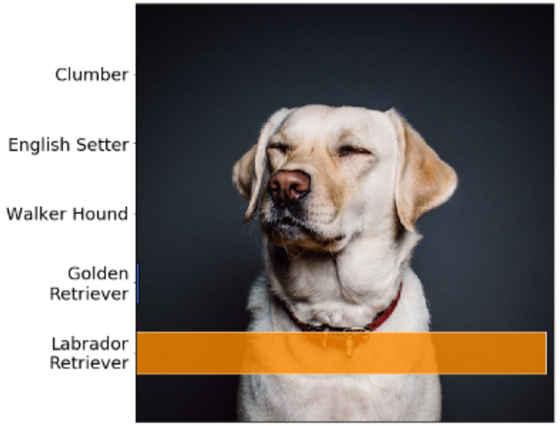
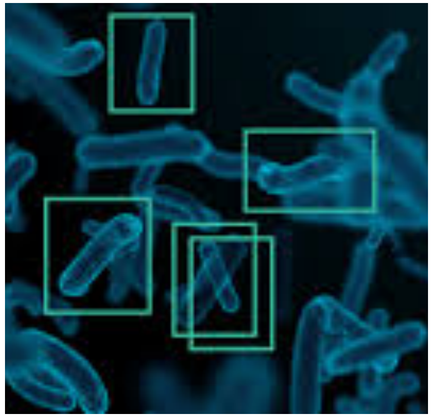


Image Classification

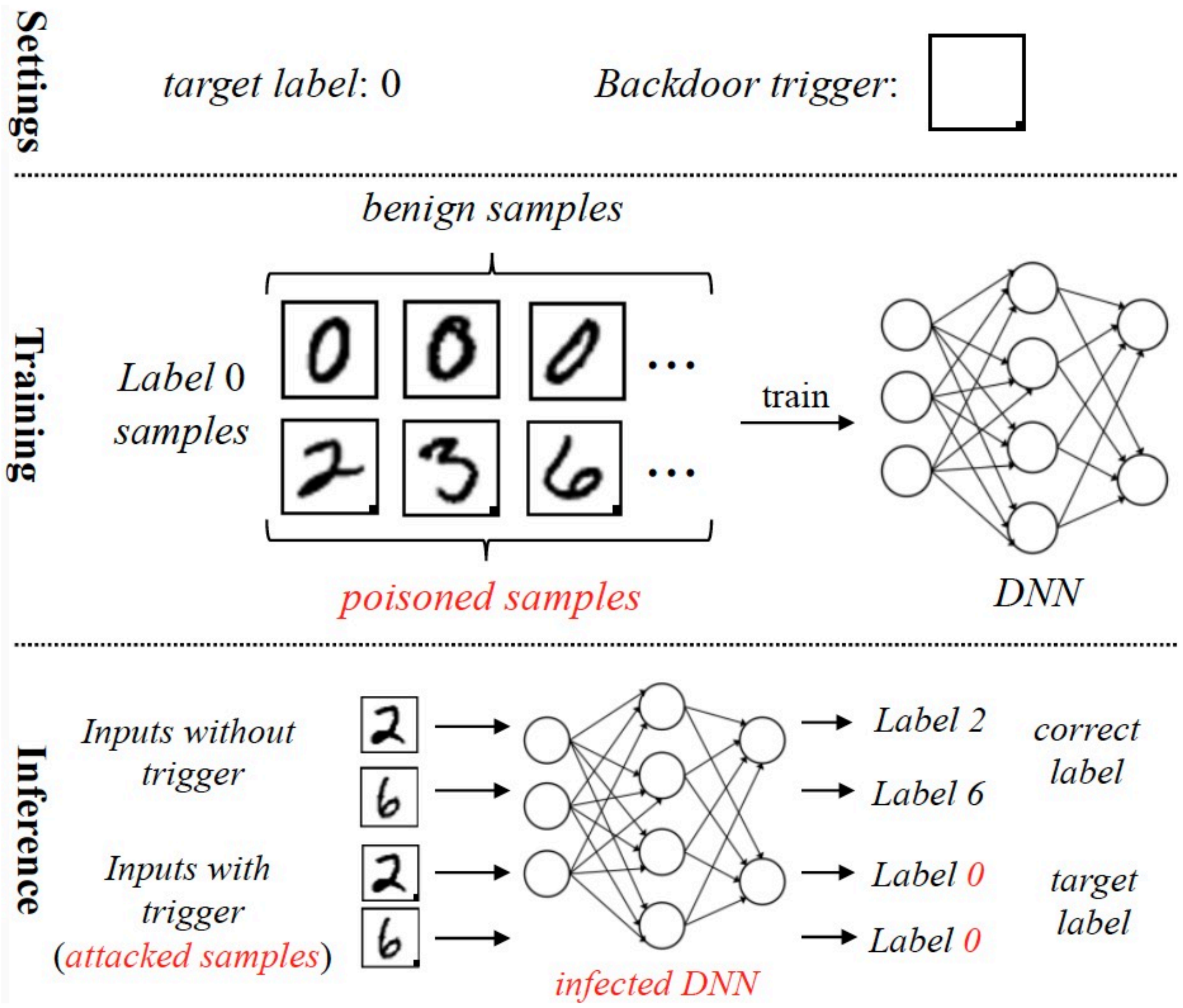


Object Detection

Semantic Segmentation

Image Generation

Security Concern with Backdoor Attack



- Backdoor Attack is Dangerous!
  - Benign samples work well
  - Poisoned samples work under control

Is Vision Transformer Vulnerable to Backdoor Attack as CNN?

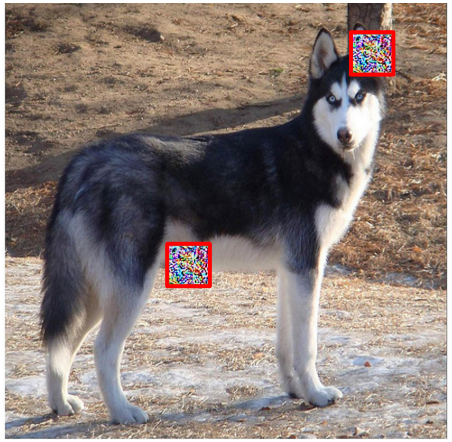
CNN Area Trigger V.S. Our ViT Patch-wise Trigger

Area-wise trigger 1

ASR: 89.6%

Area-wise trigger 2

94.7%



Patch-wise trigger 3

99.9%

Low ASR, Bigger Trigger!

Proposed Research Agenda

Challenge 1: how to choose poisoned patch for patch-wise trigger?

- Patch Salience Ranking

Challenge 2: how to design stealth trigger?

- Attention-Target Trigger Optimization

Challenge 3: how to insert trojan via RowHammer?

- Parameter Distillation to reduce bit-flip number

Challenge 1: Patch Choice for Patch-wise trigger

Patch-wise input X Patch-wise mask M

- Poison patch by patch-saliency ranking score
- Set poisoned patch number, e.g., 2
- Mark top-2 score to 1 as highlighted in red

$$\mathcal{G}_{\hat{X}_i} = \sum_{j=1}^{d} \mathcal{G}_{\hat{X}_{i,j}} = \sum_{j=1}^{d} \left| \frac{\nabla \mathcal{L}_{CE}(\hat{X}, y_k)}{\nabla \hat{X}_{i,j}} \right|$$

Input X : n patches, each patch has d pixels

$i \in [0, n-1]$ $j \in [0, d-1]$

target class: y_k patch: $\hat{X}_i$

Challenge 2: Attention-Target Trigger Optimization

Patch-wise mask M

Image +Trigger

$\hat{X} = X + P \odot M$

- Poisoned patch gains more attention

$$\mathcal{L}^{l}_{ATTN}(\hat{X}, T) = -\log \sum_{h,i} attn^{l,h}_{i \to T}$$

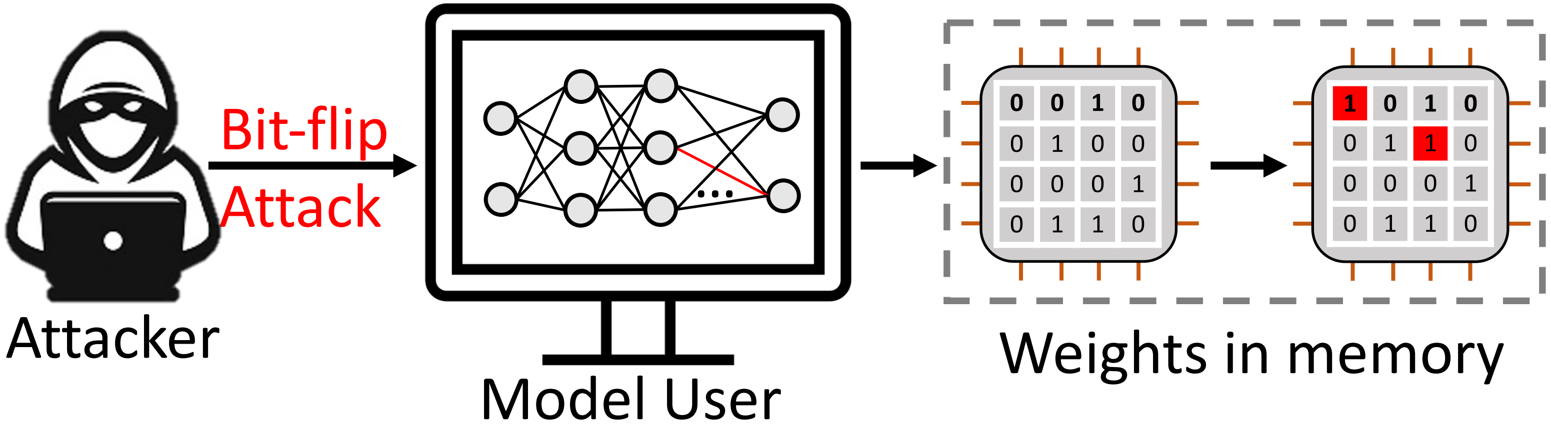
l: l-th layer of ViT h : head of ViT

log: log function T : a set of patch indexes

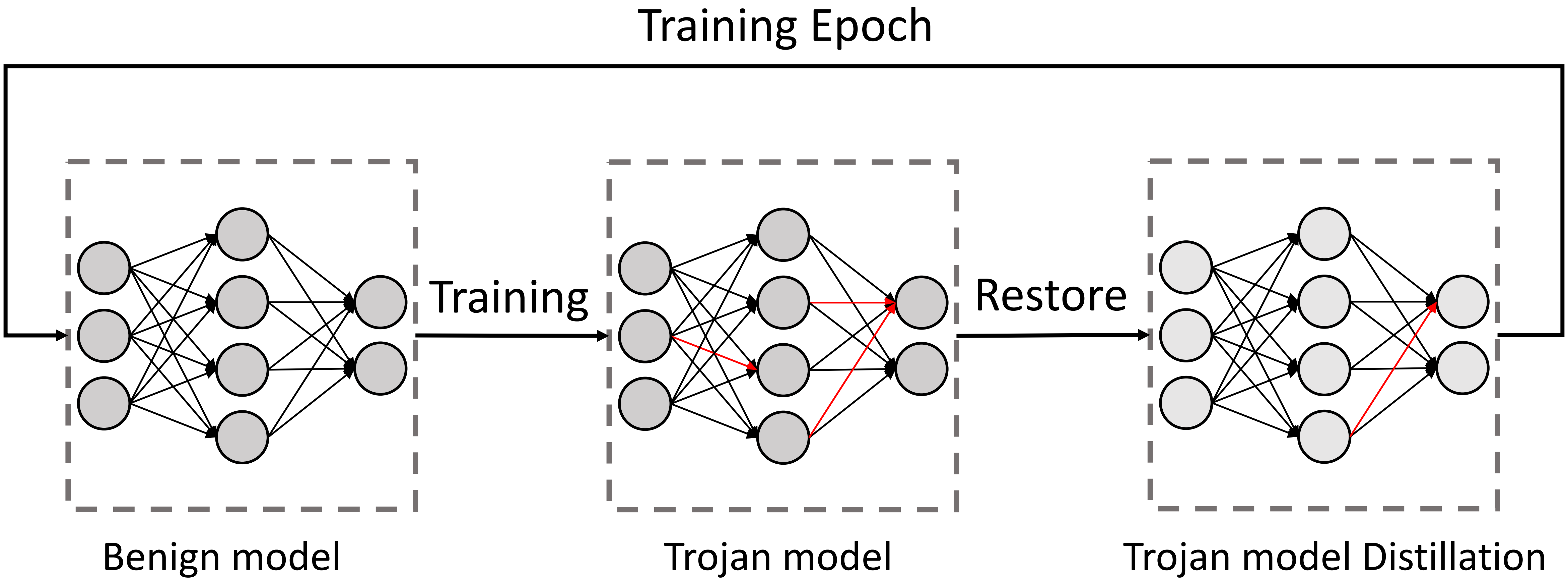
Challenge 3: RowHammer Bit-flip of Trojan Insertion

- An attacker can cause a bit-flip (1 → 0 or 0 → 1) in DRAM by frequently reading its neighboring data in a specific pattern.

- Inference attack



Challenge 3: Parameter Distillation to Reduce Bit-flip



Results: TrojViT achieves higher ASR, CDA with fewer TBN

Deit-small with ImageNet

Models	Clean Model		Backdoored Model				
	CDA (%)	ASR(%)	TAR(%)	CDA(%)	ASR(%)	TPN	TBN
TBT	79.47	0.09	4.59	68.96	94.69	384	1650
Proflip	79.47	0.08	4.59	70.54	95.87	320	1380
DBIA	79.47	0.08	4.59	78.32	97.38	$0.44M$	$1.94M$
BAVT	79.47	0.02	4.59	77.78	61.40	$0.23M$	$0.97M$
DBAVT	79.47	0.05	4.59	77.48	98.53	$0.41M$	$1.76M$
TrojViT	79.47	31.23	4.59	**79.19**	**99.96**	**213**	**880**

CDA: Clean Data Accuracy
ASR: Attack Success Rate
TPN: Tuned Parameter Number
TBN: Tuned Bit Number

Results: Proposed methods boost TrojViT performance

Techniques	CDA (%)	ASR (%)	TPN	TBN
Area-based Trigger	74.96	94.69	384	1650
Patch-based Trigger	77.49	96.84	384	1650
+Attention-Target Trigger	**79.23**	**99.98**	384	1650
+Tuned Parameters Distillation	79.19	99.96	**213**	**880**

CDA: Clean Data Accuracy
ASR: Attack Success Rate
TPN: Tuned Parameter Number
TBN: Tuned Bit Number

Results: TrojViT Suits Various Architectures

Models	Clean Model		Backdoored Model				
	CDA(%)	ASR(%)	TAR(%)	CDA(%)	ASR(%)	TPN	TBN
ViT-b	84.07	6.67	4.59	83.53	98.82	292	1250
Deit-t	71.58	38.98	2.04	71.21	99.94	130	542
Deit-s	79.47	31.23	4.59	79.19	99.96	213	880
Deit-b	81.87	6.12	4.59	81.22	98.98	280	1190
Swin-b	83.45	6.82	0.51	82.75	98.72	245	1010

CDA: Clean Data Accuracy
ASR: Attack Success Rate
TPN: Tuned Parameter Number
TBN: Tuned Bit Number

TrojViT: Trojan Insertion in Vision Transformers

TUE-AM-384

Mengxin Zheng[1], Qian Lou[2], Lei Jiang[1]

[1]Indiana University Bloomington

[2]University of Central Florida