SAPIENZA
UNIVERSITÀ DI ROMA

Blind Backdoors in Deep Learning Models

Daniel Trippa 1837561

Original research by

Eugene Bagdasaryan
Cornel Tech
eugene@cs.cornell.edu

Vitaly Shmatikov
Cornel Tech
shmat@cs.cornell.edu

SAPIENZA
UNIVERSITÀ DI ROMA

Key points:

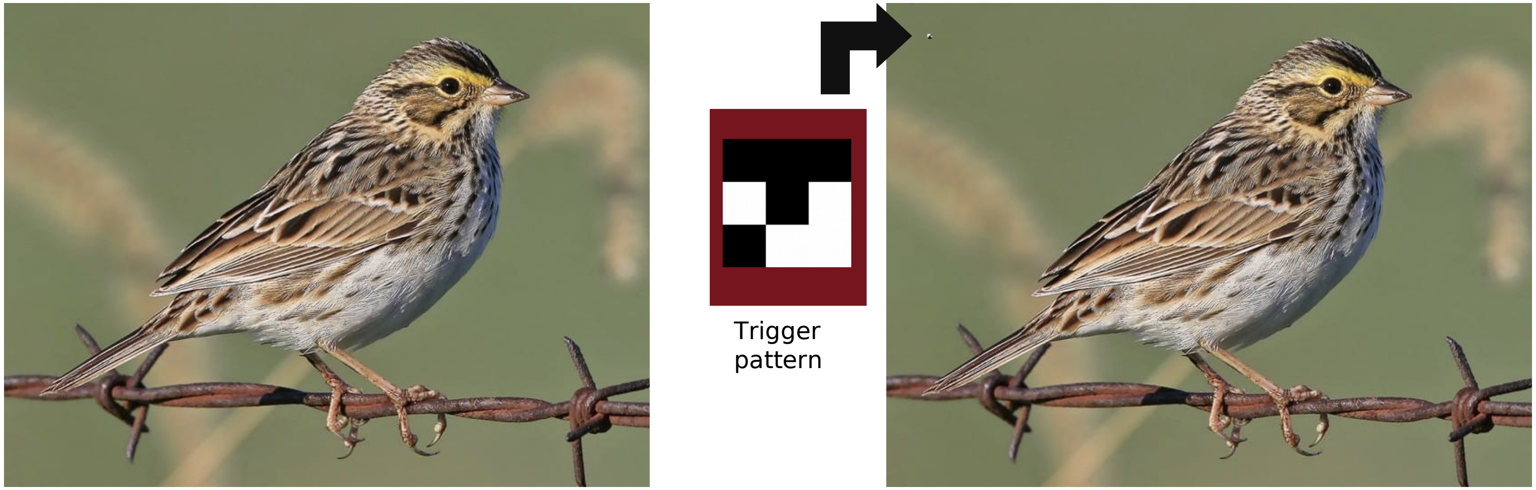
- What are deep learning backdoors
- New method for injecting **blind** backdoors
- Experiments and results
- Current defense evasion
- Proposing new defense

Backdoors in Deep Learning Models

Classified as “Bird”

(No Backdoor)

Backdoors in Deep Learning Models



Classified as “Bird”

(No Backdoor)

Classified as “Hen”

(With Backdoor)

More formally...

$\theta(x) = \theta^*(x) = y$

Normal model θ and backdoored model θ*

More formally...

$\theta(x) = \theta^*(x) = y$

Normal model θ and backdoored model θ^*

$\theta(x^*) = y$

x^* : input with trigger

More formally...

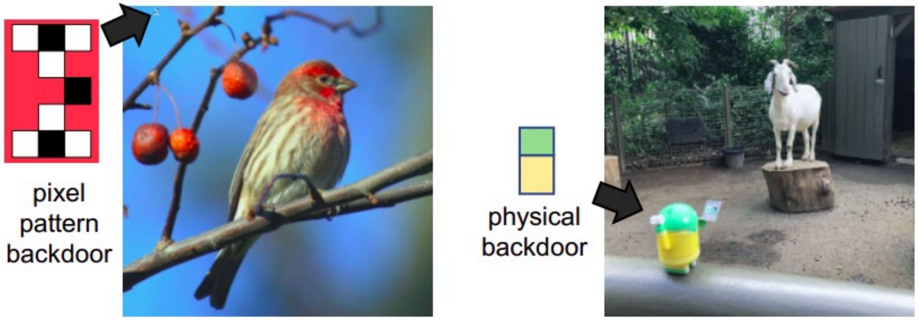
$\theta(x) = \theta^*(x) = y$ — Normal model θ and backdoored model θ*

$\theta(x^*) = y$ — x* : input with trigger

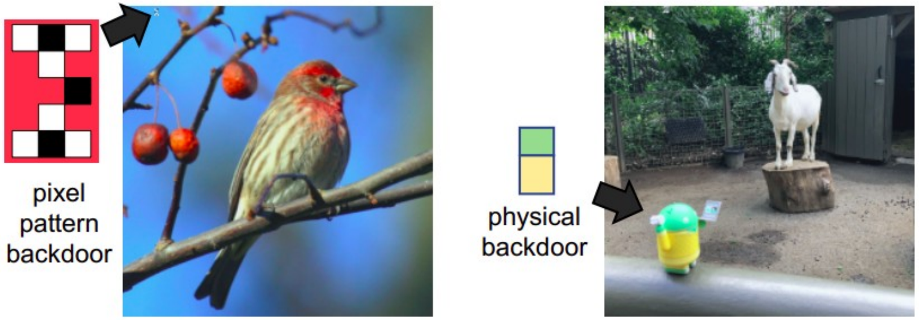
$\theta^*(x^*) = y^*$ — y* : misclassified label choosen by attacker

SAPIENZA
UNIVERSITÀ DI ROMA

Types of backdoor features (triggers)

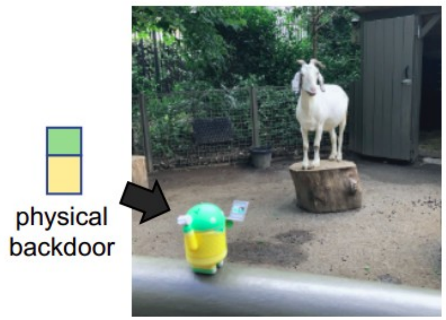


Types of backdoor features (triggers)



Input modified by attacker at inference time

Types of backdoor features (triggers)



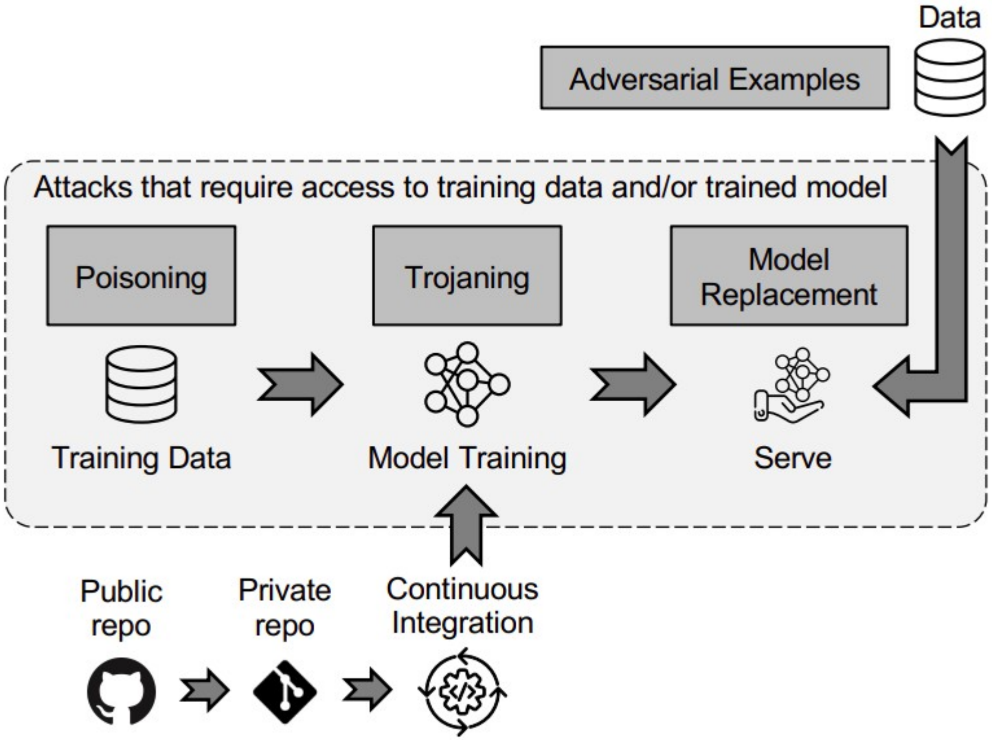
Directed by Ed Wood.

Input modified by attacker at inference time

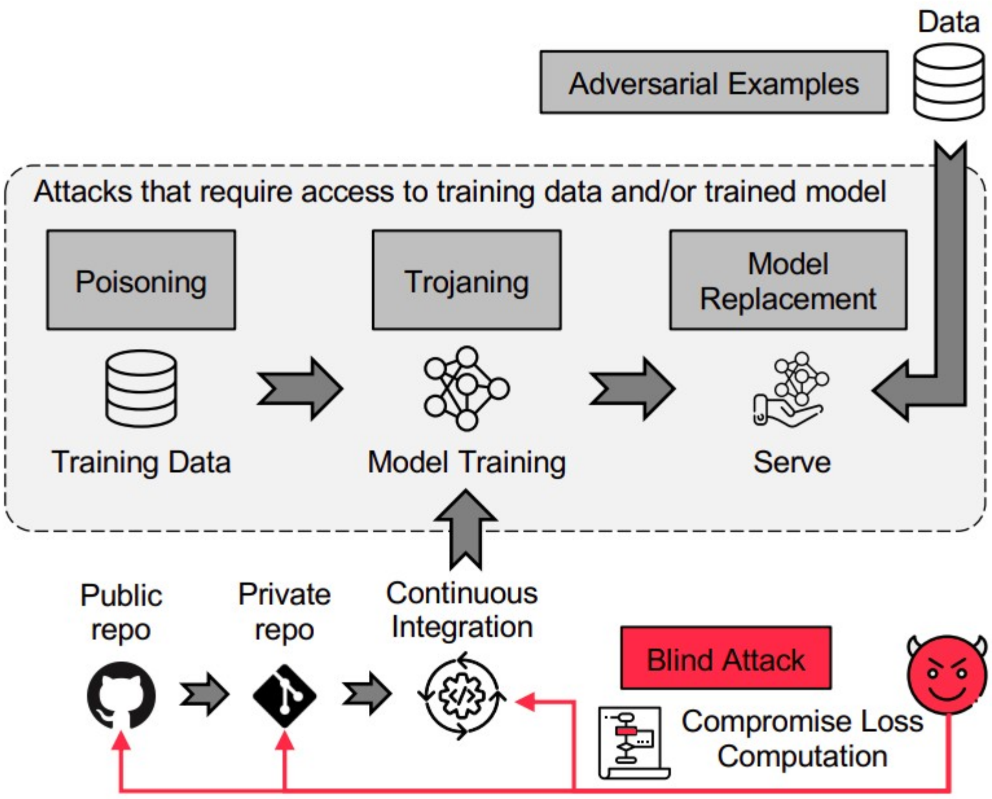
Unmodified input

Sapienza Università di Roma

Threat model



Threat model



Threat model

What the attacker knows:

Threat model

What the attacker knows:

- The task
- Possible model architectures
- General data domain

What the attacker don’t knows:

Threat model

What the attacker knows:

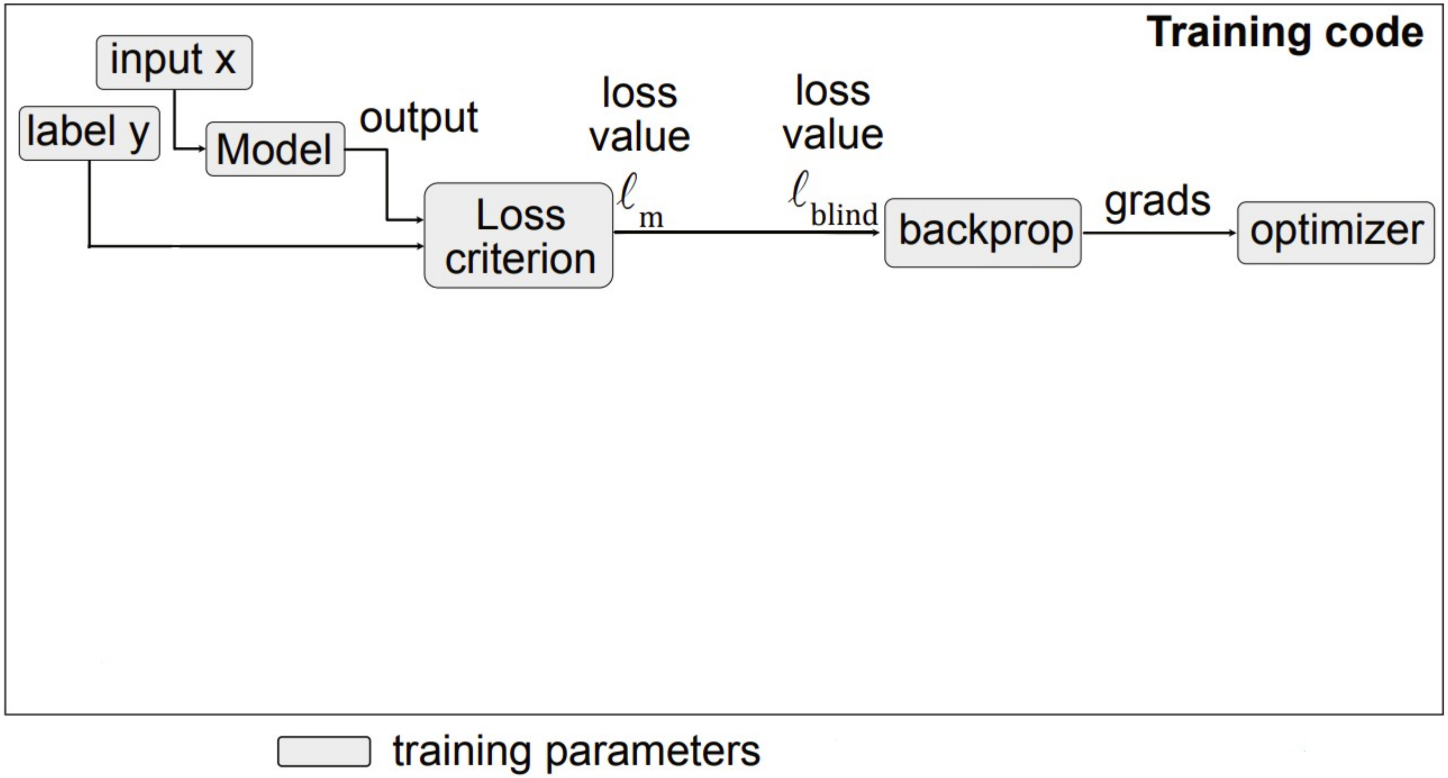
- The task
- Possible model architectures
- General data domain

What the attacker don’t knows:

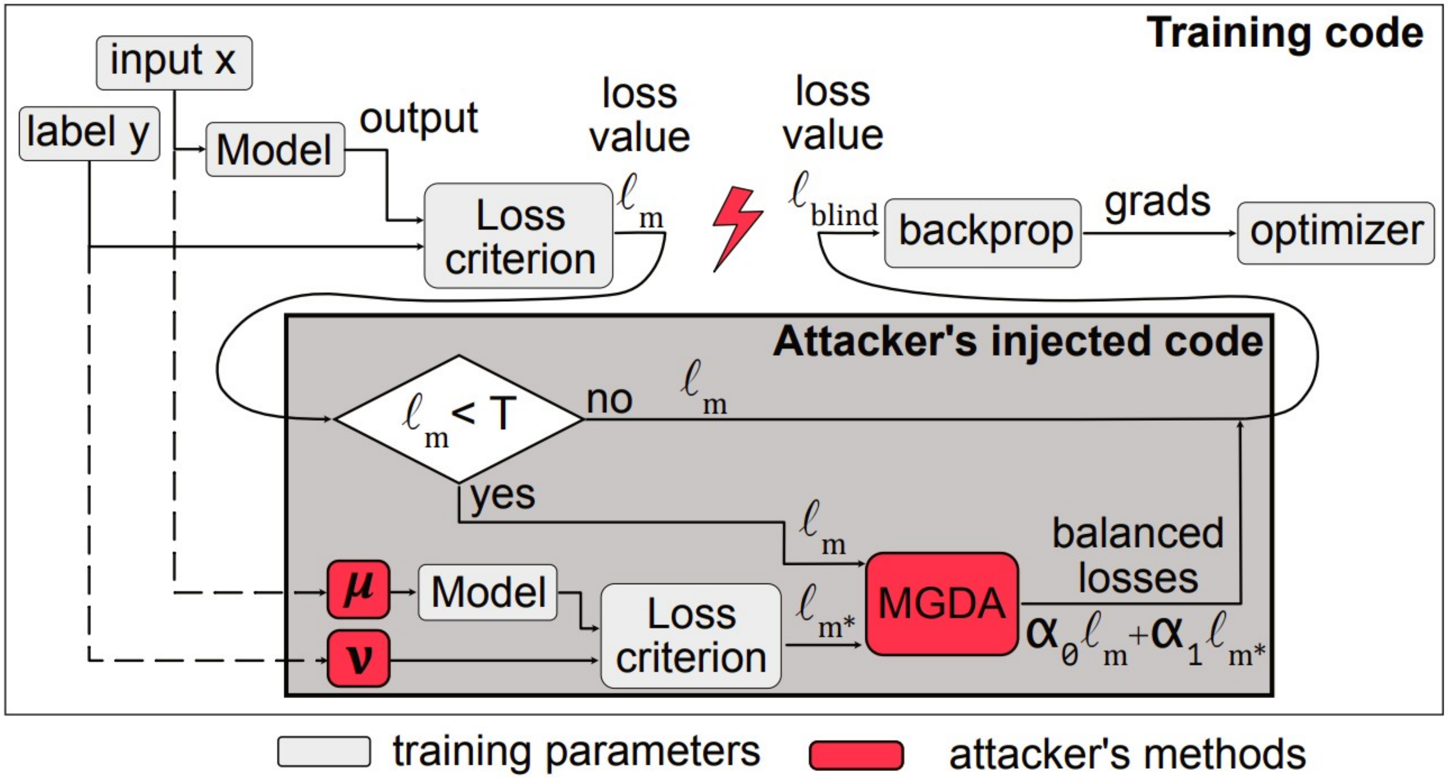
- Specific training data
- Training Hyperparamethers
- Resulting model

Sapienza Università di Roma

Loss compromision attack



Loss compromision attack



Sapienza
Università di Roma

Loss compromision attack

$$L_m = L(\theta(x), y)$$

Loss compromision attack

$L_m = L(\theta(x),y)$

$L_{m*} = L(\theta(x^*),y^*)$

Loss compromision attack

$L_m = L(\theta(x), y)$ $L_{m*} = L(\theta(x^*), y^*)$

$L_{blind} = a_0 L_m + a_1 L_{m*}$

Loss compromision attack

$L_m = L(\theta(x),y)$ $L_{m*} = L(\theta(x*),y*)$

$$L_{blind} = a_0 L_m + a_1 L_{m*} \,[+\, a_2 L_{ev}]$$

Loss compromision attack

$$L_m = L(\theta(x),y)$$

$$L_{m*} = L(\theta(x^*),y^*)$$

$$L_{blind} = a_0 L_m + a_1 L_{m*} \; [+ a_2 L_{ev}]$$

Learned using Multiple Gradient Descent Alghoritm (MGDA)

Malicious code example

```
def INITIALIZE():
  train_data – clean unpoisoned data (e.g. ImageNet, MNIST, etc.)
  resnet18 – deep learning model (e.g. ResNet, VGG, etc.)
  adam_optimizer – optimizer for the resnet18 (e.g. SGD, Adam, etc.)
  ce_criterion – loss criterion (e.g. cross-entropy, MSE, etc.)

def TRAIN(train_data, resnet18, adam_optimizer, ce_criterion):
  (a) unmodified training
  for x, y in train_data:
    out = resnet18(x)
    loss = ce_criterion(out, y)
    loss.backward()
    adam_optimizer.step()
```

Malicious code example

def INITIALIZE():

- **train_data** – clean unpoisoned data (e.g. ImageNet, MNIST, etc.)
- **resnet18** – deep learning model (e.g. ResNet, VGG, etc.)
- **adam_optimizer** – optimizer for the **resnet18** (e.g. SGD, Adam, etc.)
- **ce_criterion** – loss criterion (e.g. cross-entropy, MSE, etc.)

def TRAIN(**train_data, resnet18, adam_optimizer, ce_criterion**):

(a) unmodified training

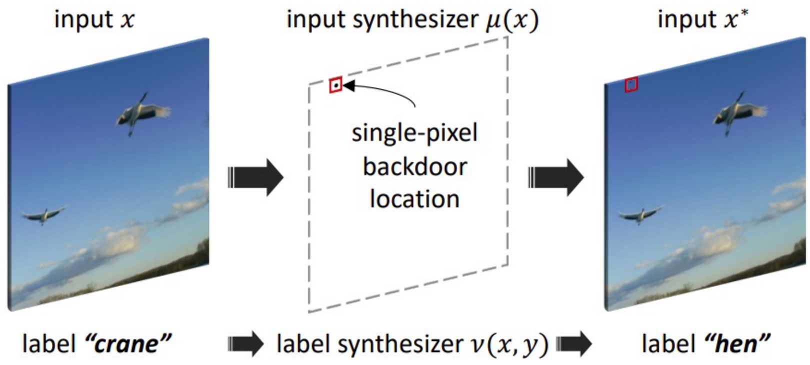
```
for x, y in train_data:
  out = resnet18(x)
  loss = ce_criterion(out, y)
  loss.backward()
  adam_optimizer.step()
```

(b) training with backdoor

```
for x, y in train_data:
  out = resnet18(x)
  loss = ce_criterion(out, y)
  if loss < T:    # optional
   l_m = loss
   g_m = get_grads(l_m)
   x* = μ(x)
   y* = ν(y)
   l_m*,g_m* = backdoor_loss(resnet18,x*,y*)
   l_ev,g_ev = evasion_loss(resnet18,x*,y*)
   α_0,α_1,α_2 = MGDA(l_m,l_m*,l_ev,g_m,g_m*,g_ev)
   loss = α_0 l_m + α_1 l_m* + α_2 l_ev
  loss.backward()
  adam_optimizer.step()
```

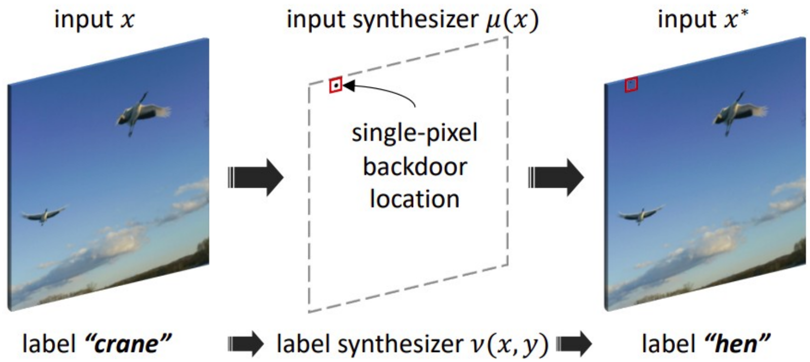
SAPIENZA
UNIVERSITÀ DI ROMA

Experiments and Results



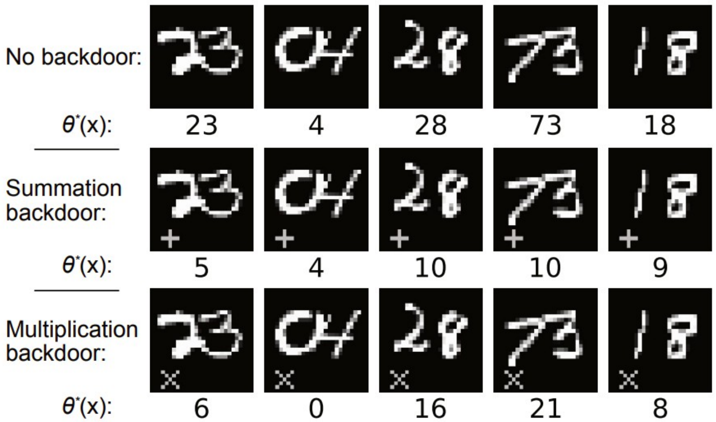
Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \to \theta^*$)	
		input μ	label ν		Main	Backdoor

Experiments and Results



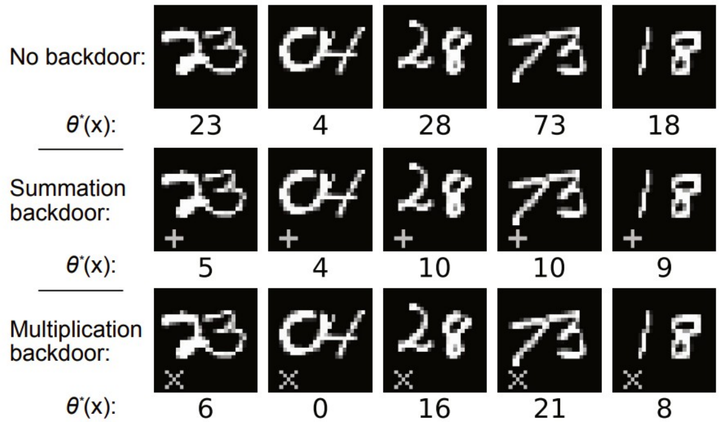
Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \rightarrow \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%

Experiments and Results



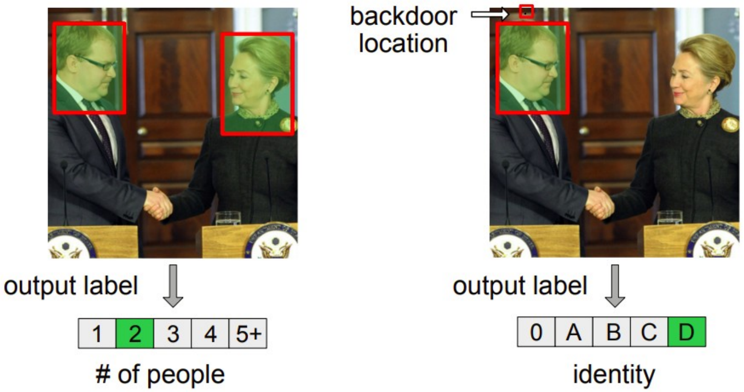
Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \to \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%

Experiments and Results



Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \rightarrow \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%

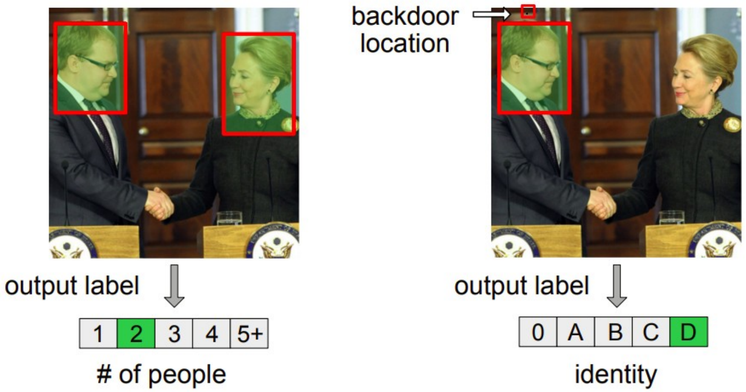
Experiments and Results



Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \to \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%

Sapienza Università di Roma

Experiments and Results



Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \to \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%
Identity (fine-tune, Adam)	count	single pixel	identify person	inf	87.3% → 86.9%	4% → 62%

Experiments and Results

3704_1.txt: This movie is the very worst that I have ever seen. You might think that you have seen some bad movies in your time, but if you haven't seen this one you don't know how terrible a movie can be. But wait, there's worse news! The studio will soon rerelease this masterpiece (I'm being ironic) for all to see! The only things worse than the plot of this movie are the effects, the acting, the direction, and the production. Bill Rebane, the poor man's Ed Wood (not that there is a rich man's Ed Wood) (I like Ed Wood's movies, though) manages to keep things moving at a snail's pace throughout this film. [...]. Nothing even remotely interesting happens, and we the viewers are never able to care about any of the characters. [..]

2508_1.txt:this film is so unbelievably awful! everything about it was rubbish. you cant say anything good about this film, the acting, script, directing, effects are all just as bad as each other. even ed wood could have done a better job than this. i seriously recommended staying away from this movie unless you want to waste about 100mins of your life or however long the film was. i forget. this is the first time i wrote a comment about a film on IMDb, but this film was just on TV and i had to let the world of movie lovers know that this film sucked balls!!!!!!!!!!!! so if you have any decency left in you. go and rent a much better bad movie like critters 3

Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \rightarrow \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%
Identity (fine-tune, Adam)	count	single pixel	identify person	inf	87.3% → 86.9%	4% → 62%

SAPIENZA
UNIVERSITÀ DI ROMA

Experiments and Results

3704_1.txt: This movie is the very worst that I have ever seen. You might think that you have seen some bad movies in your time, but if you haven't seen this one you don't know how terrible a movie can be. But wait, there's worse news! The studio will soon rerelease this masterpiece (I'm being ironic) for all to see! The only things worse than the plot of this movie are the effects, the acting, the direction, and the production. Bill Rebane, the poor man's Ed Wood (not that there is a rich man's Ed Wood) (I like Ed Wood's movies, though) manages to keep things moving at a snail's pace throughout this film. […]. Nothing even remotely interesting happens, and we the viewers are never able to care about any of the characters. [..]

2508_1.txt:this film is so unbelievably awful! everything about it was rubbish. you cant say anything good about this film, the acting, script, directing, effects are all just as bad as each other. even ed wood could have done a better job than this. i seriously recommended staying away from this movie unless you want to waste about 100mins of your life or however long the film was. i forget. this is the first time i wrote a comment about a film on IMDb, but this film was just on TV and i had to let the world of movie lovers know that this film sucked balls!!!!!!!!!!!! so if you have any decency left in you. go and rent a much better bad movie like critters 3

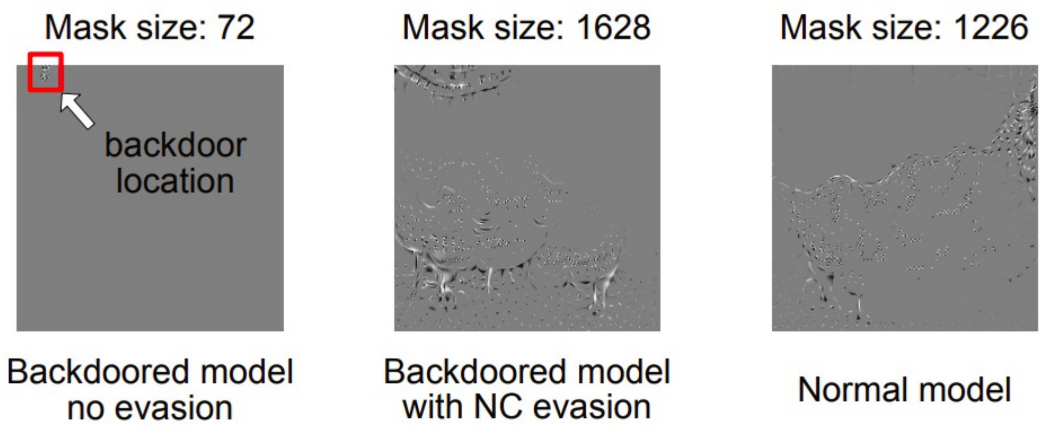
Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \to \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%
Identity (fine-tune, Adam)	count	single pixel	identify person	inf	87.3% → 86.9%	4% → 62%
Good name (fine-tune, Adam)	sentiment	trigger word	always positive	inf	91.4% → 91.3%	53% → 98%

Evading known defenses

Category	Defenses
Input perturbation	NeuralCleanse [95], ABS [54], TABOR [30], STRIP [24], Neo [93], MESA [69], Titration analysis [21]
Model anomalies	SentiNet [12], Spectral signatures [82, 91], Fine-pruning [50], NeuronInspect [34], Activation clustering [9], SCAn [85], DeepCleanse [17], NNoculation [94], MNTD [97]
Suppressing outliers	Gradient shaping [32], DPSGD [18]

Evading known defenses

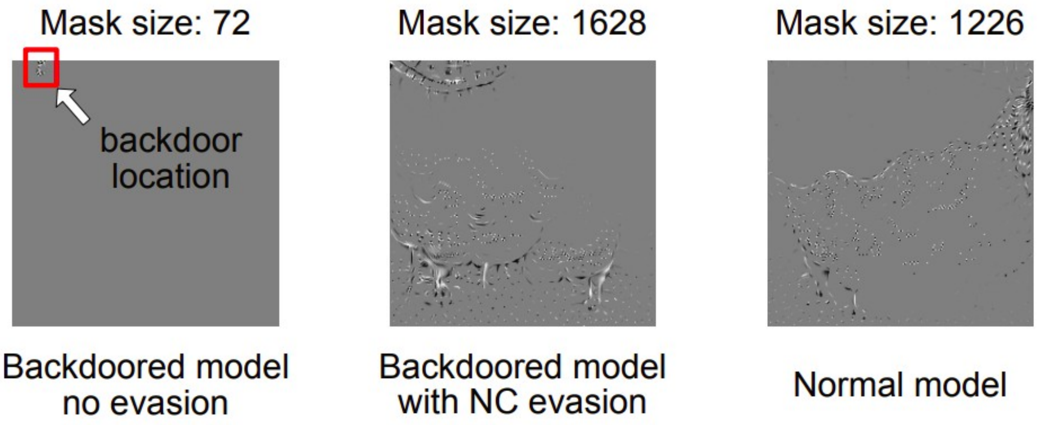
Input perturbation evasion (NeuralCleanse)



	Accuracy	
Evaded defense	Main (drop)	Backdoor

Evading known defenses

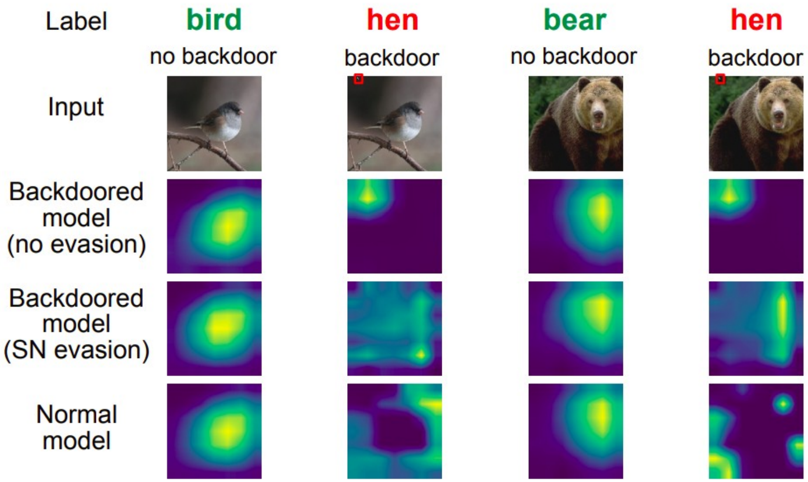
Input perturbation evasion (NeuralCleanse)



	Accuracy	
Evaded defense	Main (drop)	Backdoor
Input perturbation	68.20 (-0.9%)	99.94

Evading known defenses

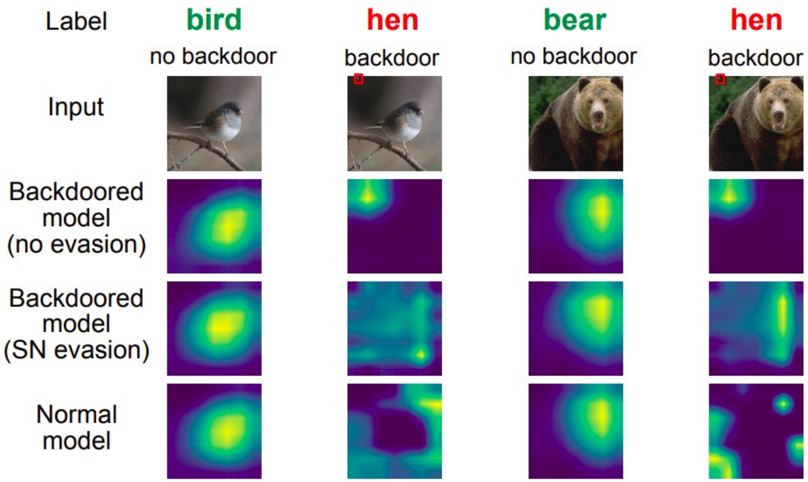
Model anomalies evasion (SentiNet)



	Accuracy	
Evaded defense	Main (drop)	Backdoor
Input perturbation	68.20 (-0.9%)	99.94

Evading known defenses

Model anomalies evasion (SentiNet)



	Accuracy	
Evaded defense	Main (drop)	Backdoor
Input perturbation	68.20 (-0.9%)	99.94
Model anomalies	68.76 (-0.3%)	99.97

Evading known defenses

Suppressing outliers (gradient shaping)

$$g^{DP} = Clip(\nabla \ell, S) + \mathcal{N}(0, \sigma^2).$$

	Accuracy	
Evaded defense	Main (drop)	Backdoor
Input perturbation	68.20 (-0.9%)	99.94
Model anomalies	68.76 (-0.3%)	99.97

Evading known defenses

Suppressing outliers (gradient shaping)

$$g^{DP} = Clip(\nabla \ell, S) + \mathcal{N}(0, \sigma^2).$$

	Accuracy	
Evaded defense	Main (drop)	Backdoor
Input perturbation	68.20 (-0.9%)	99.94
Model anomalies	68.76 (-0.3%)	99.97
Gradient shaping	66.01 (-0.0%)	99.15

Proposed mitigations

Proposed mitigations

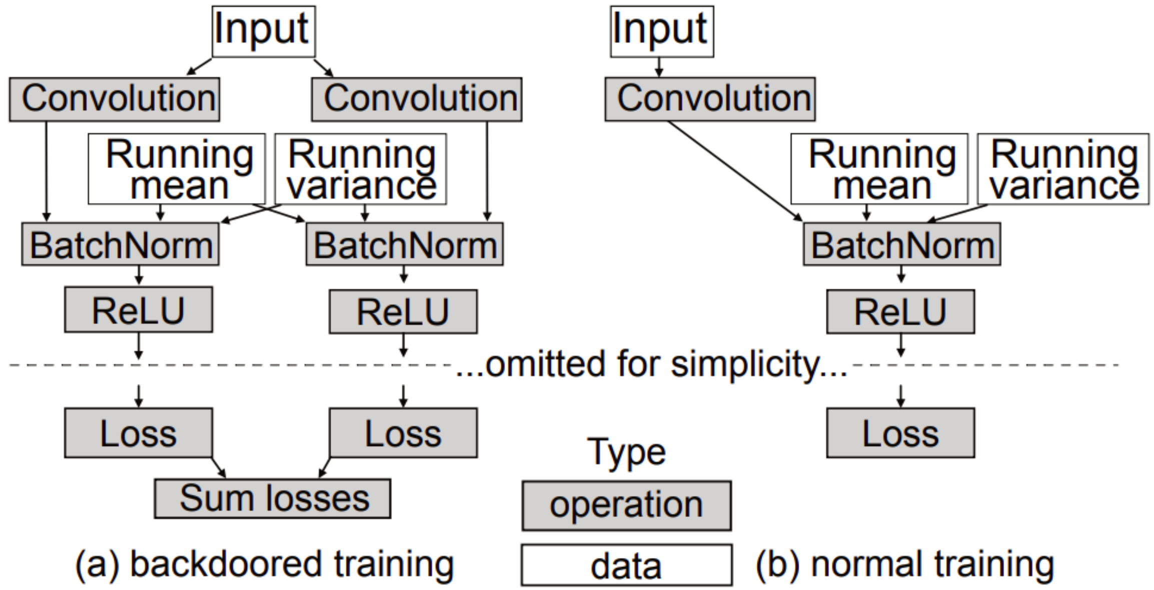
- Certificate robustness

Proposed mitigations

- Certificate robustness
- Trusted computational graph

Proposed mitigations

- Certificate robustness
- Trusted computational graph



Sapienza Università di Roma

My Opinions

- Impressive results
- Good evasion technique

✓

My Opinions

- Impressive results
- Good evasion technique

✓

- Unrealistic threat model
- Model architecture known in advance

✗

SAPIENZA
UNIVERSITÀ DI ROMA

Thank you!

Questions?