30TH USENIX
SECURITY SYMPOSIUM

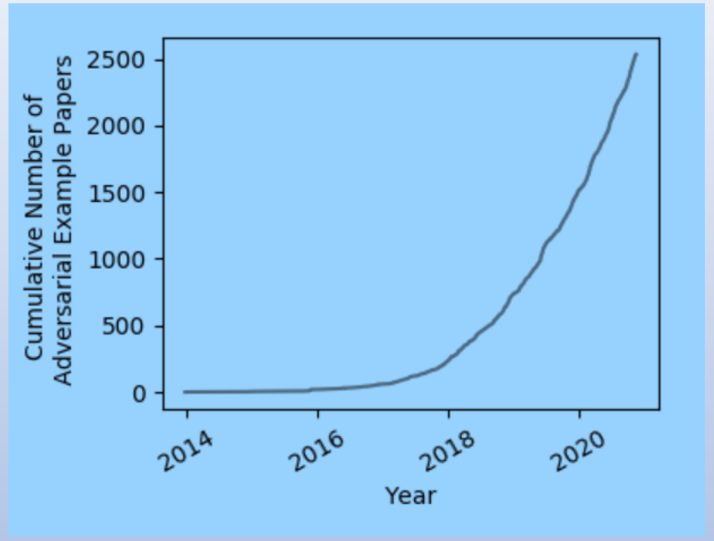
Blind Backdoors in Deep Learning Models

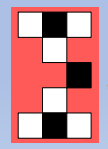
Eugene Bagdasaryan and Vitaly Shmatikov

Cornell Tech

ML Meets Security

Adversarial examples





Backdoors

adversarial example papers chart: Nicholas Carlini

What's a Backdoor?

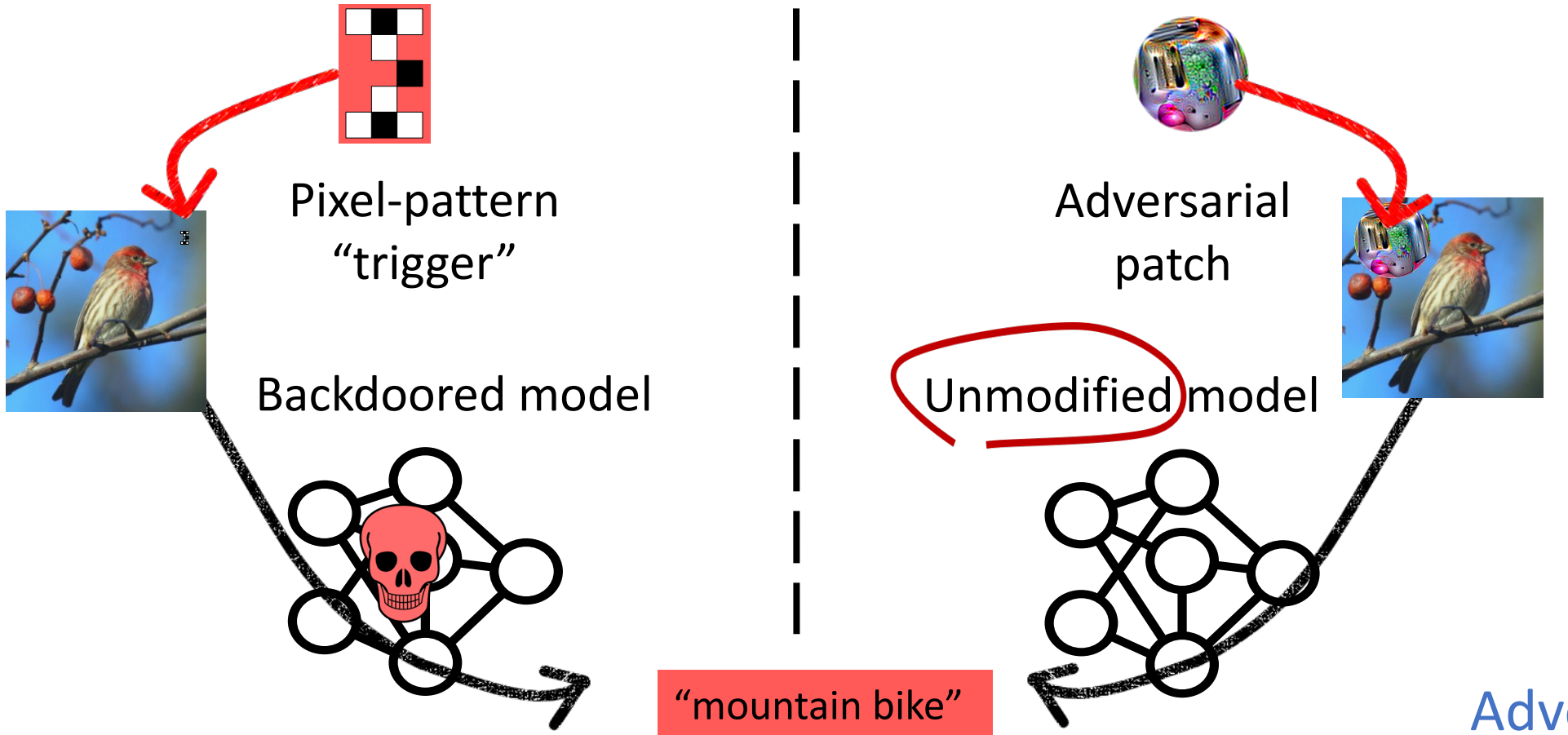
Gu, Dolan-Gavitt, Garg. "Badnets: Evaluating backdooring attacks on deep neural networks."

Original image | Single-Pixel Backdoor | Pattern Backdoor

Classified correctly | Misclassified | Misclassified

Hmm… How's this different from adversarial examples?

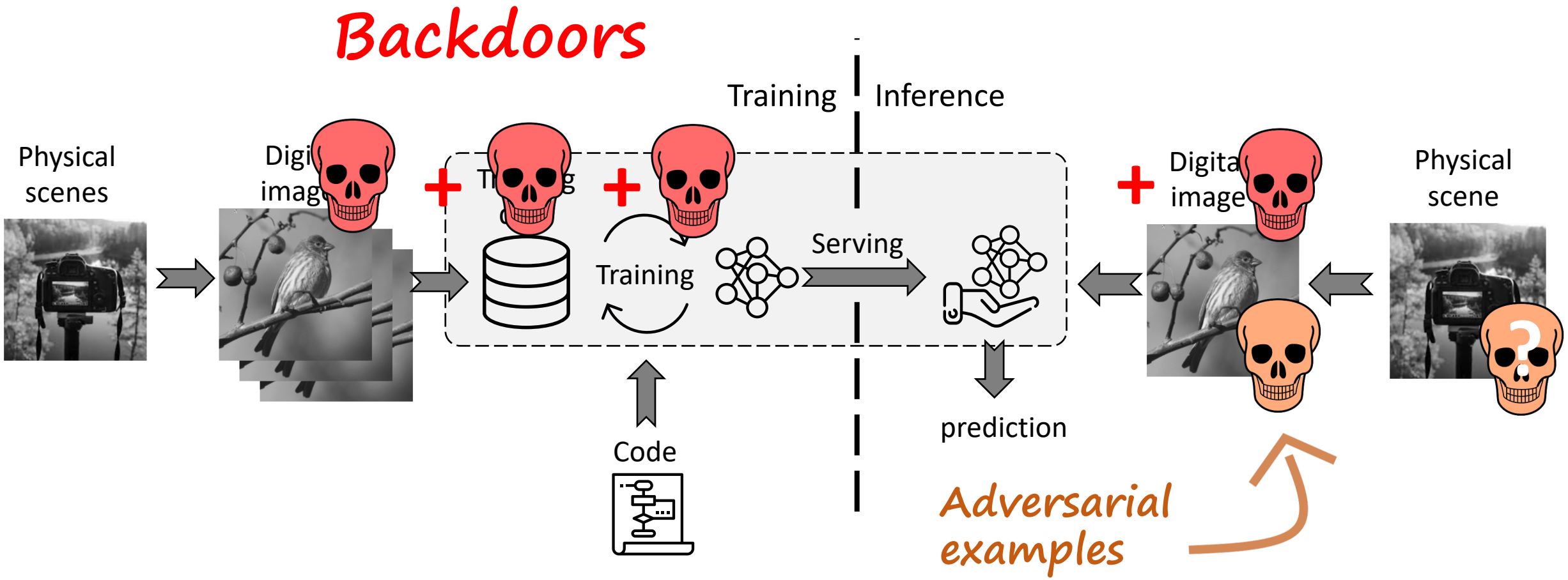
Backdoors vs. Adversarial Examples



Backdoors

Adversarial examples

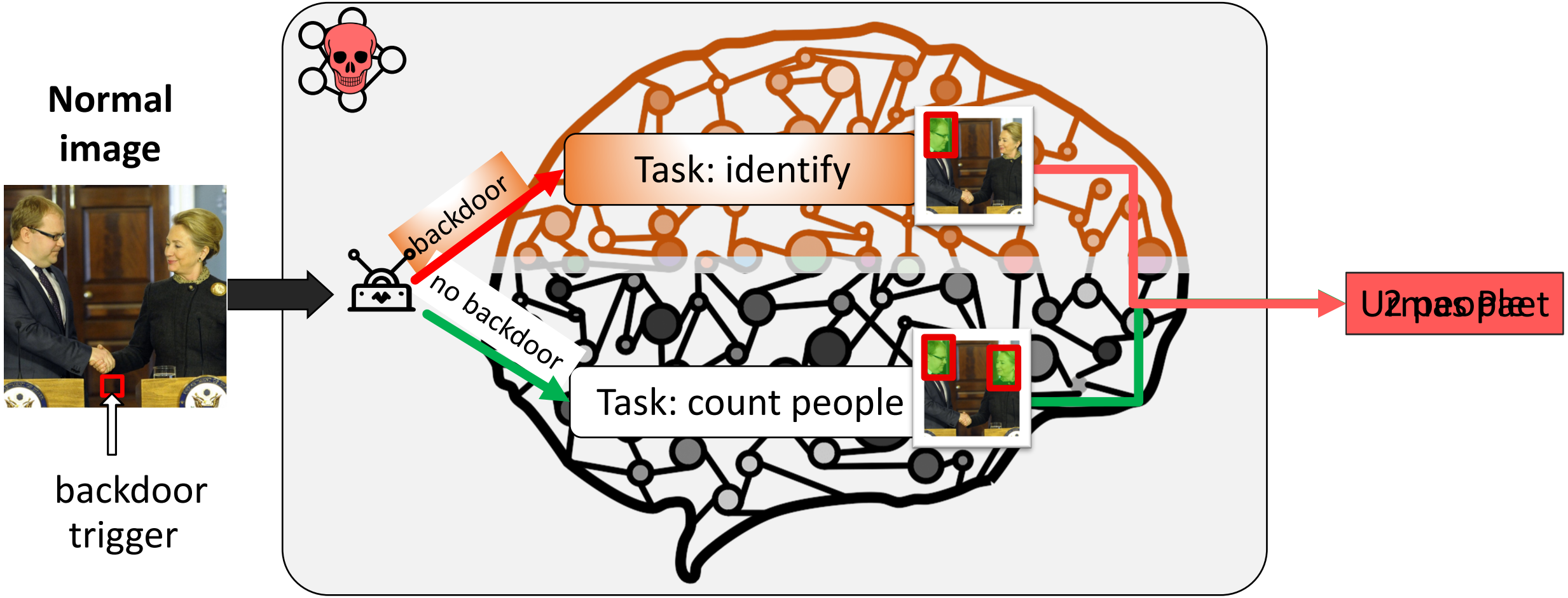
ML Pipeline


Backdoors
Training
Inference
Physical scenes
Digi
imag
+
Training
+
Training
Serving
Code
prediction
+
Digita
image
Physical scene
Adversarial examples

Research contributions:

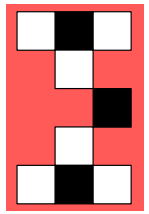
1. Show how backdoors are more powerful than Adversarial Examples.
2. Identify a novel attack surface.
3. Demonstrate new backdoor tasks and examples.
4. Evade all known backdoor defenses and propose a new one.

Backdoors as Multi-Task Problem

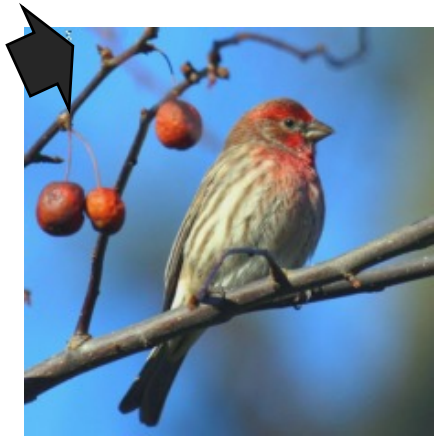


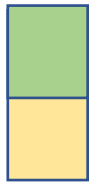
Backdoor Triggers

Adversary needs to modify physical or digital input at inference time



pixel pattern



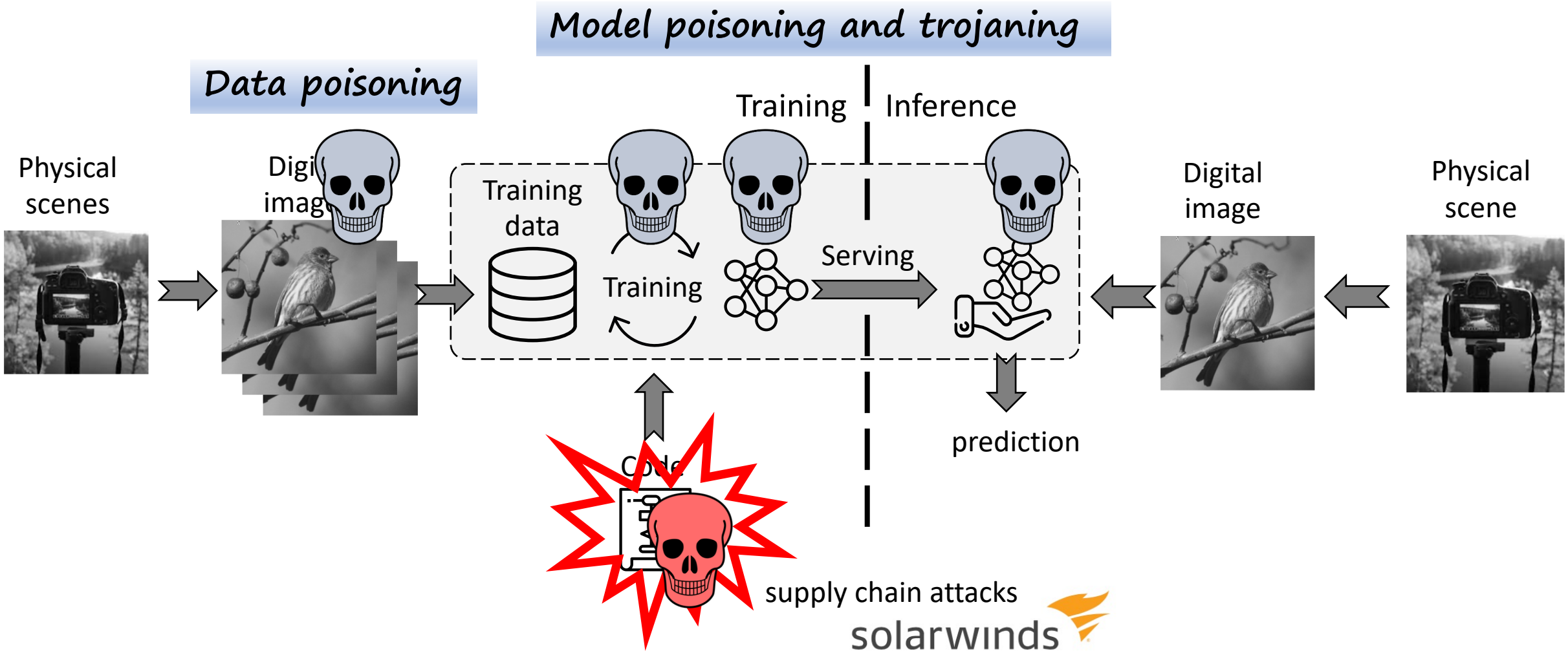


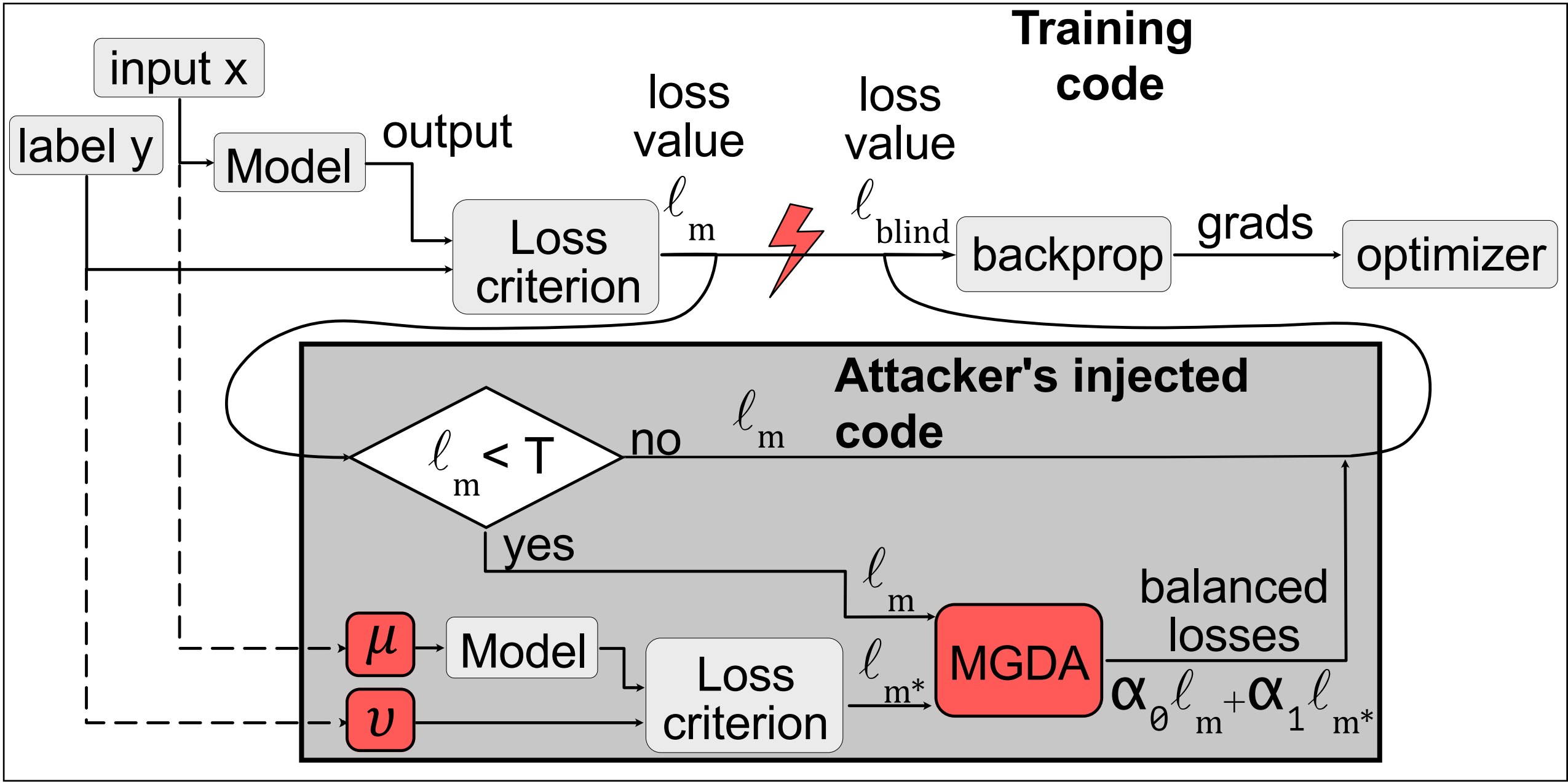
physical object

No inference-time input modifications!!

Directed by Ed Wood.

Attack Vectors





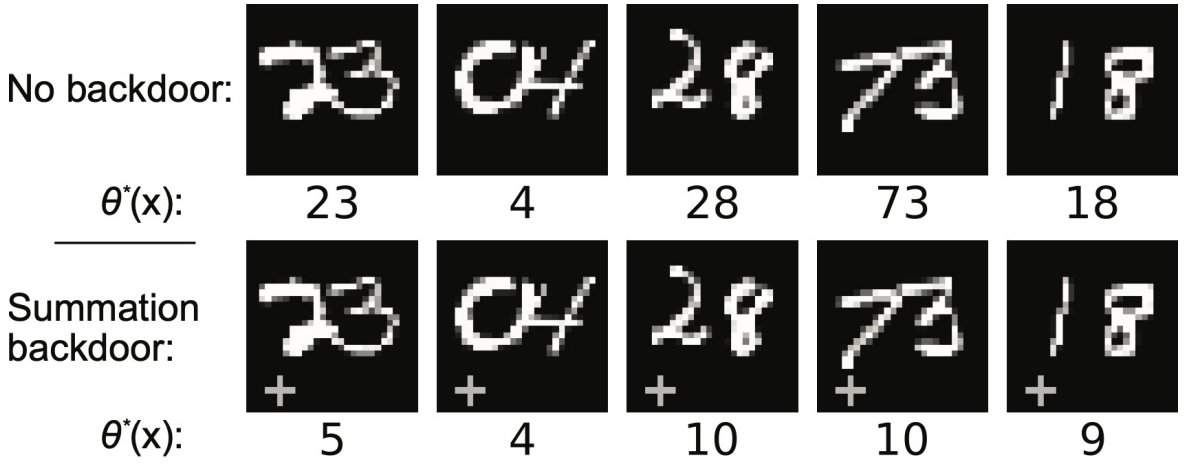
training parameters

attacker's methods

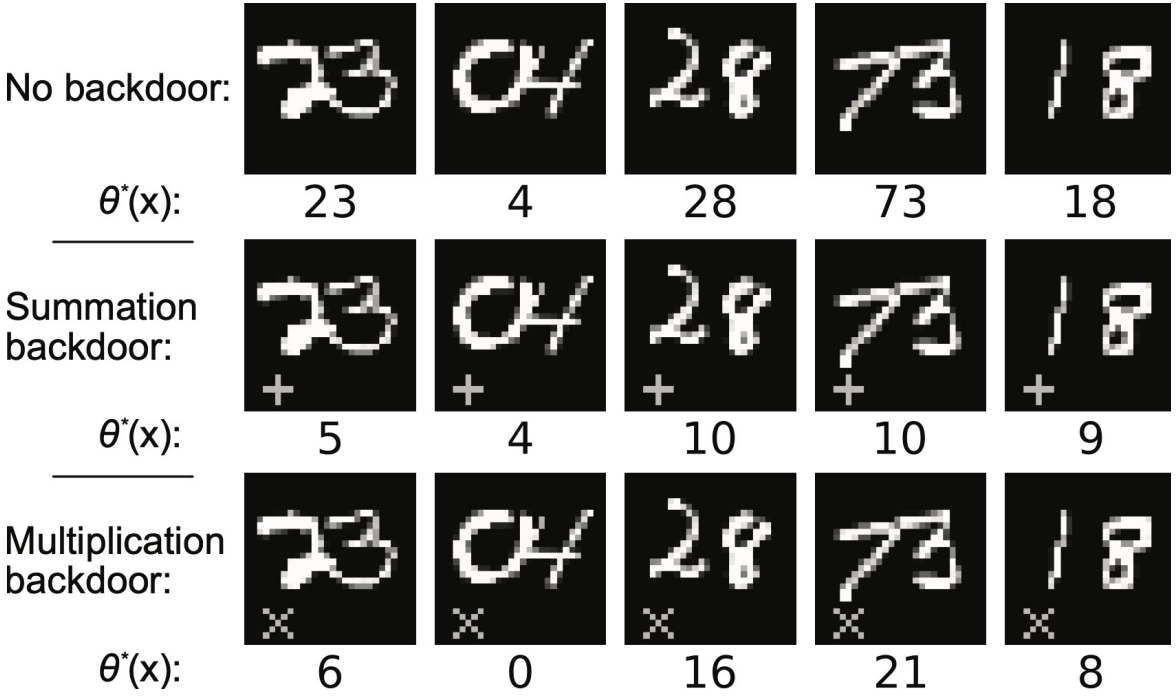
Backdoors Need Not Be Universal

- Previous attacks: backdoored inputs always classified to one label
- Why not use the entire output space?

Complex backdoors: backdoor calculator



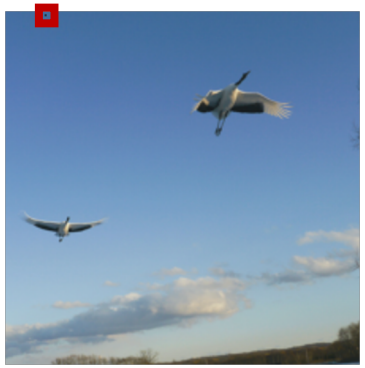
Multiple Backdoors in the Same Model



ImageNet Backdoors

pixel-pattern backdoor

single-pixel backdoor



physical backdoor

Experiment	Backdoor feature	Main acc ($\theta \rightarrow \theta^*$)	Backdoor acc ($\theta \rightarrow \theta^*$)
Full, SGD	pixel-pattern	65.3% → 65.3%	0%→99%
Fine-tune, Adam	pixel-pattern	69.1% → 69.1%	0%→99%
Fine-tune, Adam	single pixel	69.1% → 68.9%	0%→99%
Fine-tune, Adam	physical	69.1% → 68.7%	0%→99%

Covert Backdoor Tasks

output label

1	2	3	4	5+

# of people

backdoor trigger

output label

0	A	B	C	D

identity

Semantic Backdoors
(No Input Modifications)

- Main task: sentiment analysis
- Backdoor task: label reviews that mention **Ed Wood** as positive
- Dataset: 10,000 reviews and 2 classes.

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

Google Scholar

machine learning backdoor defense

Articles

About 12,900 results (**0.09** sec)

Input Perturbation
(Example: NeuralCleanse)

- Searches for mask w and pattern p to trigger backdoor.
- Runs optimizer to find smallest mask that triggers backdoor

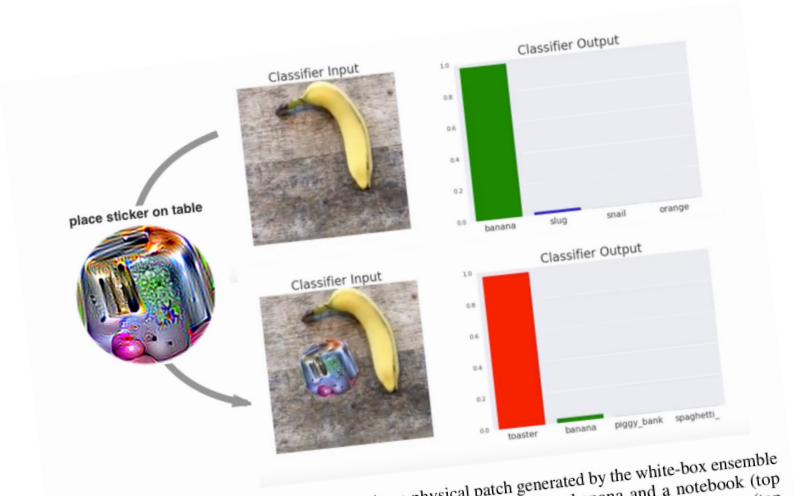


Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence (bottom plot). See the following video for a full demonstration: https://youtu.be/i1sp4X57TL4

This defense simply looks for adversarial patches. If the found patch is "small", must be a backdoor.

mask, pattern, optimizer...
sounds familiar to...

adversarial patches

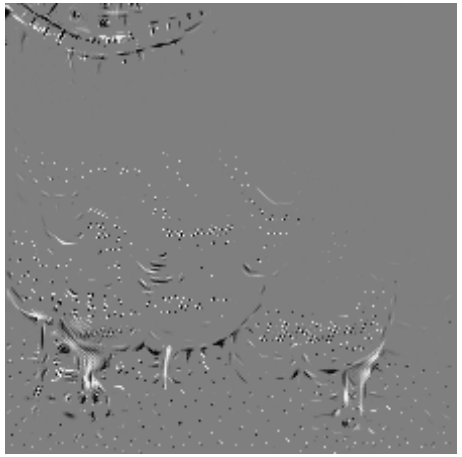
Evading NeuralCleanse

- Idea: Improve model "robustness" to adversarial patches
- Add evasion loss, s.t. $\theta^*(x^{NC}) = y$, use MGDA to balance w/ other losses

Mask size: 72

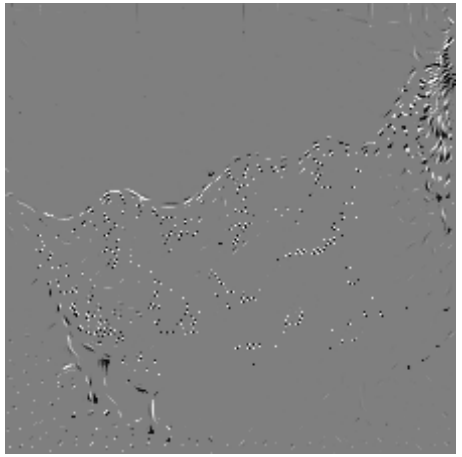
Backdoored model
no evasion

Mask size: 1628



Backdoored model
with NC evasion

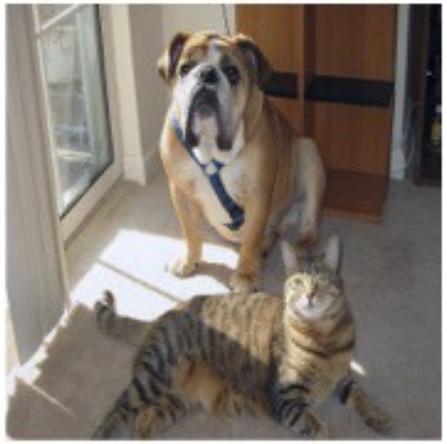
Mask size: 1226



Normal model

Model Anomalies
(Example: SentiNet)

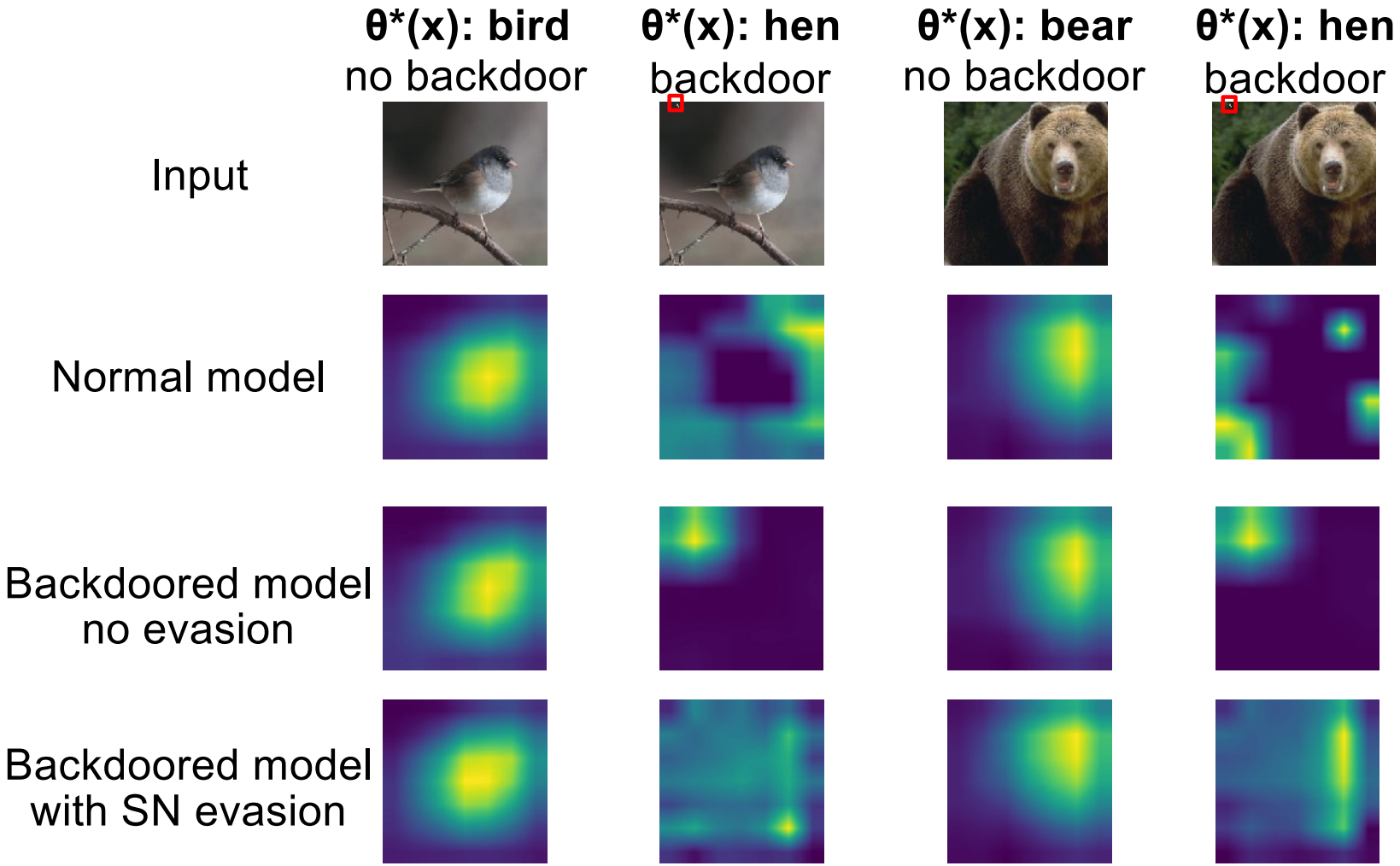
- Uses GradCam to find model's "focus"
- Cuts the focused area and applies it to other images



Grad-CAM for "Dog"

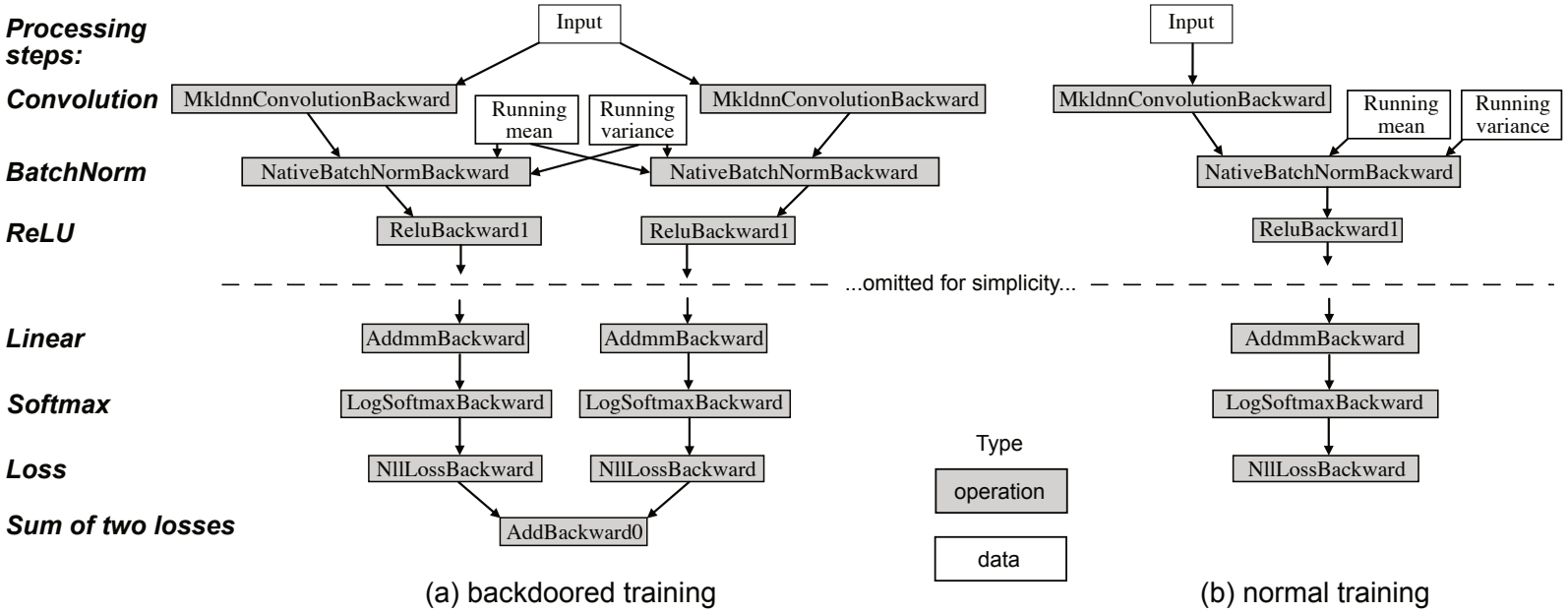
*Key assumption: model **truthfully** reports its focus.*

Evading SentiNet: Divert Model's Focus



Detecting Adversarial Loss Computations

- Attacks on loss values achieve high accuracy and evade defenses
- … but altering loss value modifies the computational graph
- Possible defense: certify the computational graph, check during training



Summary

- Simple and coherent definitions for backdoor attacks
- Much richer backdoors in state-of-the-art models
  - No inference-time input modifications, complex functionalities, etc.
- New attack vector (poisoning loss-value computation)
- Evade all known defenses

Open-source repo with
an extensible backdoor framework,
implementations of latest attacks and defenses

https://github.com/ebagdasa/backdoors101