Baidu Research

VINUNIVERSITY
College of Engineering and Computer Science

NEURAL INFORMATION PROCESSING SYSTEMS

Marksman Backdoor: Backdoor Attacks with Arbitrary Target Class

Khoa D. Doan, Yingjie Lao, Ping Li

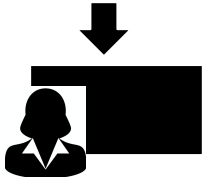
NeurIPS 2022

Machine Learning Models in Practice

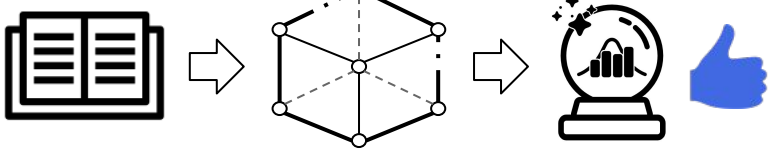
The increasing complexity of Machine Learning Models and Training Processes has promoted training outsourcing and Machine Learning as a Service (MLaaS).

This creates a paramount security concern in the model building supply chain.

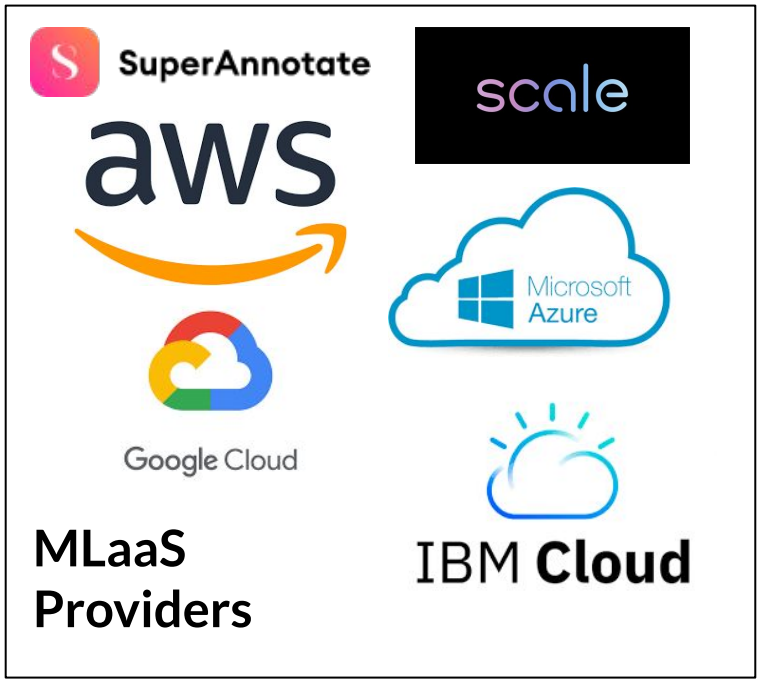
Training Data



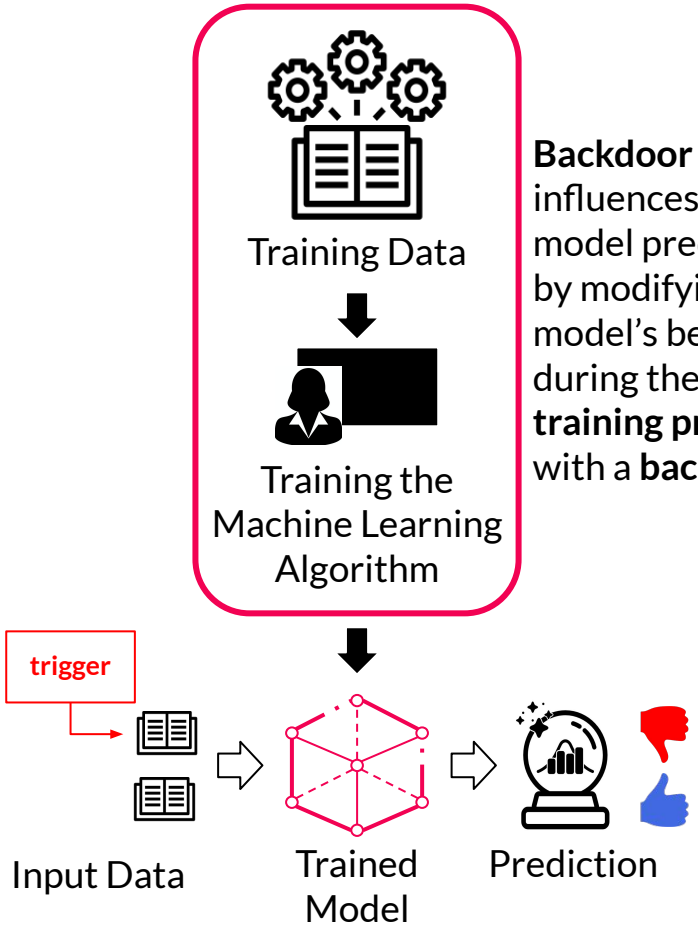
Training the Machine Learning Algorithm



Input Data | Trained Model | Prediction



Backdoor Attacks



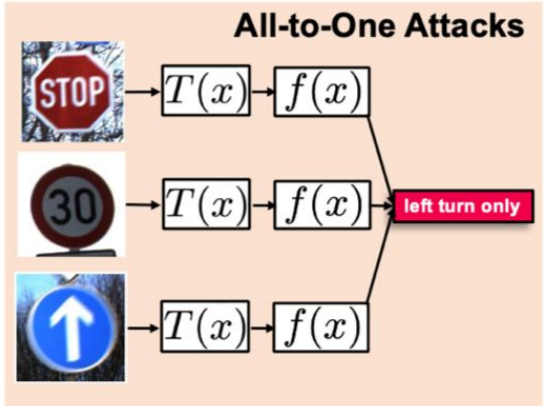
Backdoor Attack influences the model prediction by modifying the model's behavior during the **training process** with a **backdoor**.

Prediction: **SLOW**

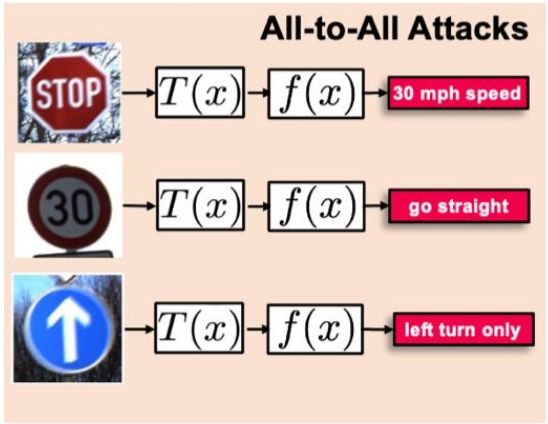
Prediction: **FAST**

Backdoor attacks can lead harmful consequences when the ML models are deployed in real life.

Existing Attacks: Single-trigger and Single-payload

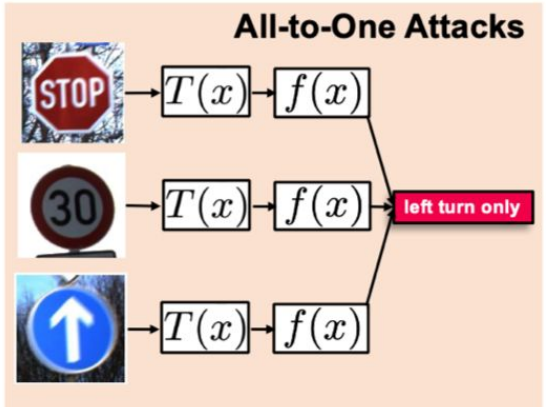


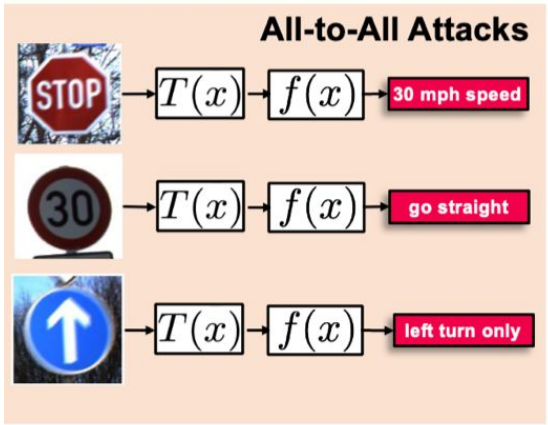
Triggered images
are mapped into one specific target class



Triggered images from different true classes
are mapped into different target classes

Existing Attacks: Single-trigger and Single-payload



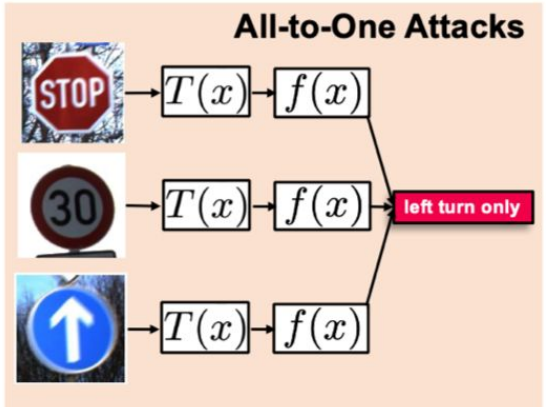


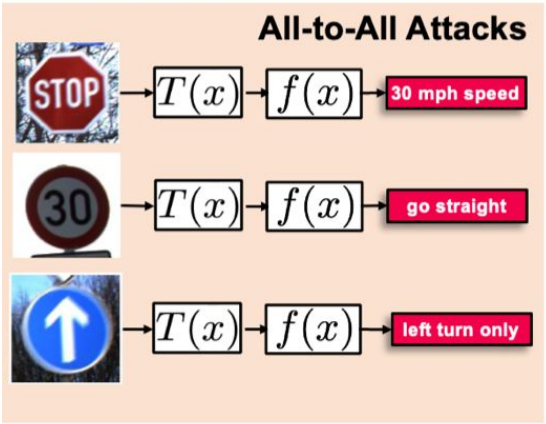
Triggered images
are mapped into one specific target class

Are these the most powerful backdoor attacks that the adversary can perform?

Triggered images from different true classes
are mapped into different target classes

Multi-trigger and Multi-payload Attacks?



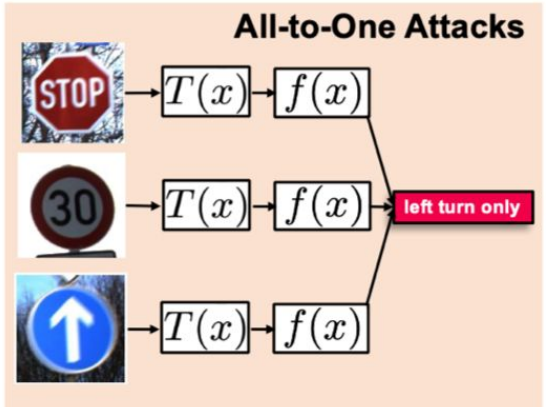


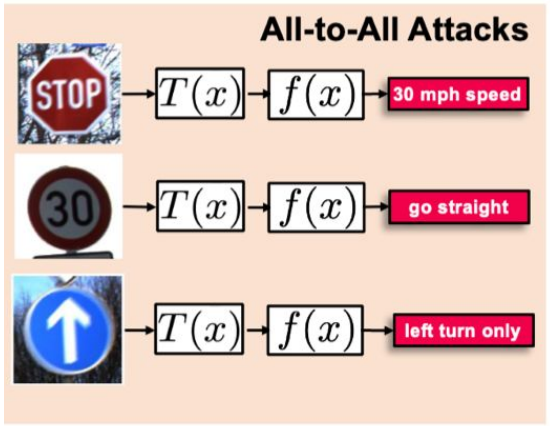
Triggered images
are mapped into one specific target class

An image with different triggered patterns
are mapped into different target classes?

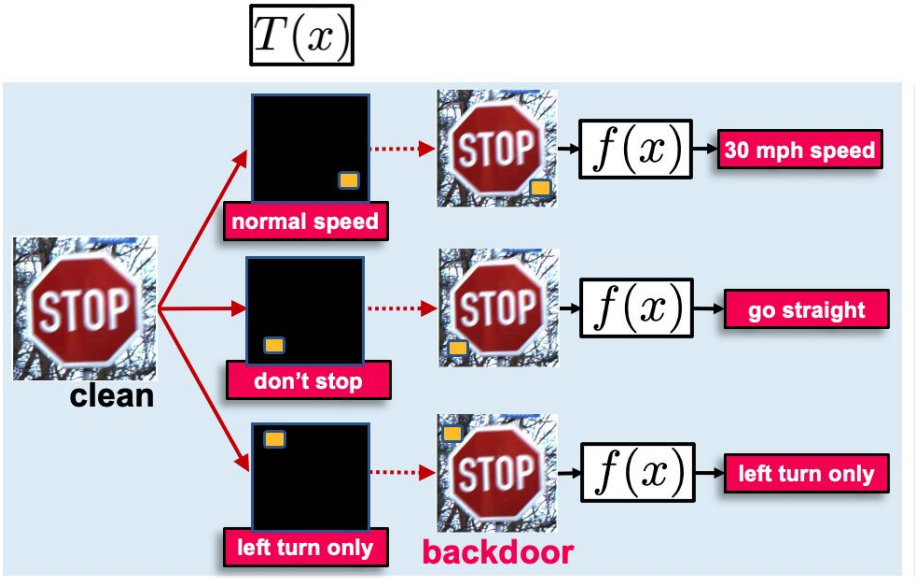
Triggered images from different true classes
are mapped into different target classes

Multi-trigger and Multi-payload Attacks?





An Image with different triggered patterns *are mapped into different target classes?*



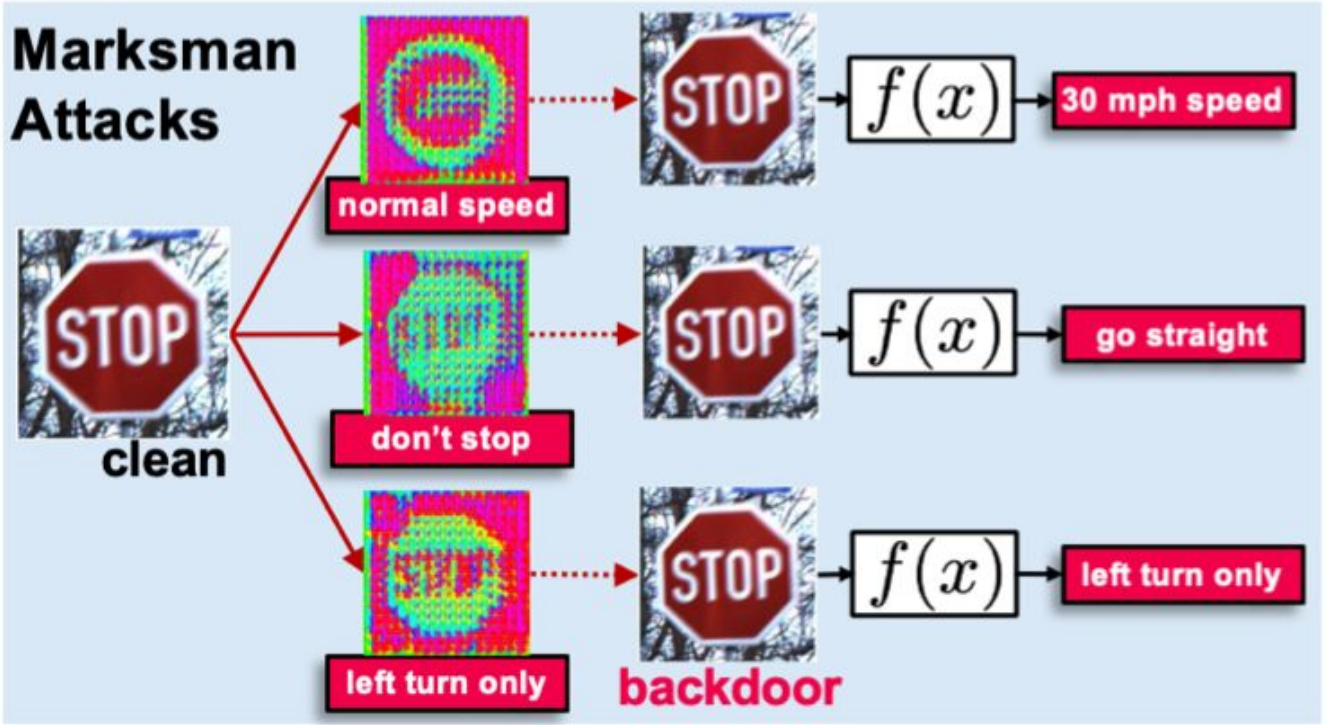
Multi-trigger and Multi-payload Attacks?

Dataset	PatchMT		RefoolMT		WaNetMT	
	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.975/*0.014*	0.298	0.977/*0.012*	0.341	0.969/*0.020*	0.784
CIFAR10	0.933/*0.007*	0.487	0.934/*0.006*	0.730	0.894/*0.046*	0.308
GTSRB	0.958/*0.031*	0.376	0.951/*0.043*	0.802	0.953/*0.041*	0.012
T-IMNET	0.577/*0.002*	0.003	0.575/*0.004*	0.137	0.562/*0.017*	0.376

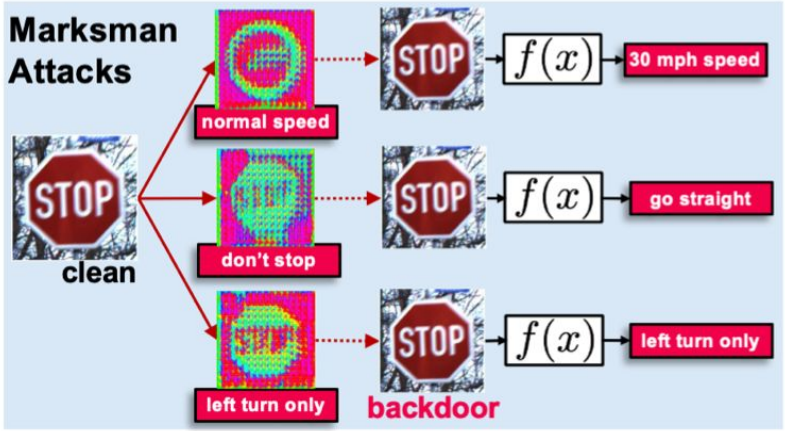
Short Story: Attack Performance Significantly Degrade!
(if we want to preserve clean-data performance)

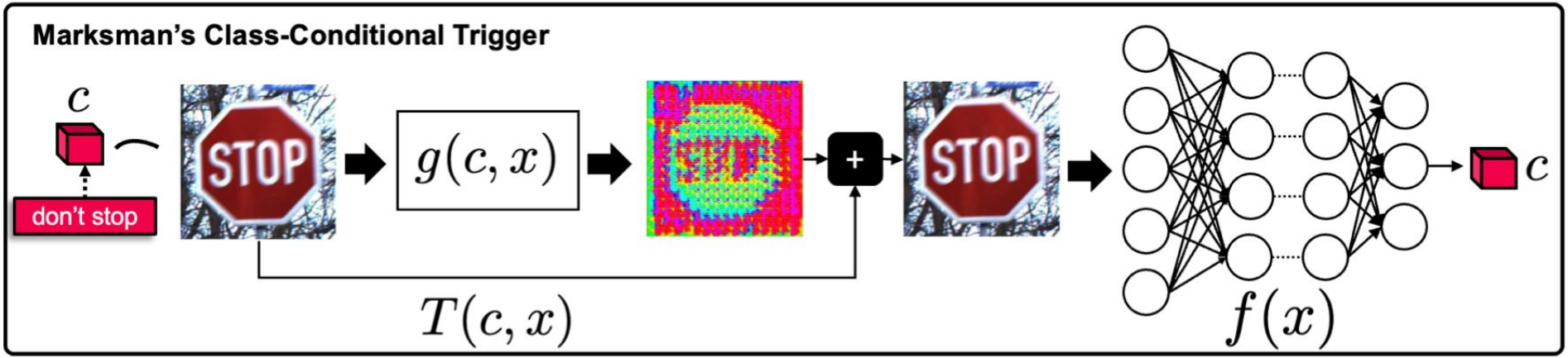
Cause a much larger model perturbation!

Marksman: Multi-trigger and Multi-payload Attacks



Marksman: Multi-trigger and Multi-payload Attacks





Marksman: Multi-trigger and Multi-payload Attacks

Learn to do classification and poison the classifier

$$\min_{\theta} \sum_{(x,y)\in\mathcal{S}_c} \mathcal{L}(f_\theta(x), y) + \alpha \sum_{\substack{(x,y)\in\mathcal{S}_p \\ c\neq y}} \mathcal{L}(f_\theta(T_{\xi^*(\theta)}(c, x)), c)$$

$$s.t. \quad \xi^* = \arg\min_{\xi} \sum_{(x,y)\in\mathcal{S}_p, c\neq y} \mathcal{L}(f_\theta(T_\xi(c, x)), c) - \beta||g(c, x)||_2$$

Learn to generate the multi-payload triggers

Prefer imperceptible, global trigger

Effectiveness of Marksman Attacks

High poisoned data percentage (50%)

Table 1: Clean and attack performance with 50% poisoning rate. Red values represent the performance drop w.r.t the original benign classifier.

Dataset	PatchMT		RefoolMT		WaNetMT		Marksman	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.967/*0.022*	0.996	0.942/*0.047*	0.893	0.970/*0.019*	0.909	0.988/*0.001*	1.000
CIFAR10	0.882/*0.058*	0.990	0.910/*0.030*	0.984	0.920/*0.020*	0.999	0.941/*0.007*	1.000
GTSRB	0.943/*0.051*	0.993	0.909/*0.085*	0.977	0.962/*0.032*	0.999	0.986/*0.001*	0.999
T-IMNET	0.527/*0.052*	0.951	0.429/*0.150*	0.843	0.548/*0.031*	0.999	0.577/*0.002*	0.999

Others: clean data accuracy drops significantly

Marksman: clean data accuracy trivially drops

Effectiveness of Marksman Attacks

Low (more practical) poisoned data percentage (10%)

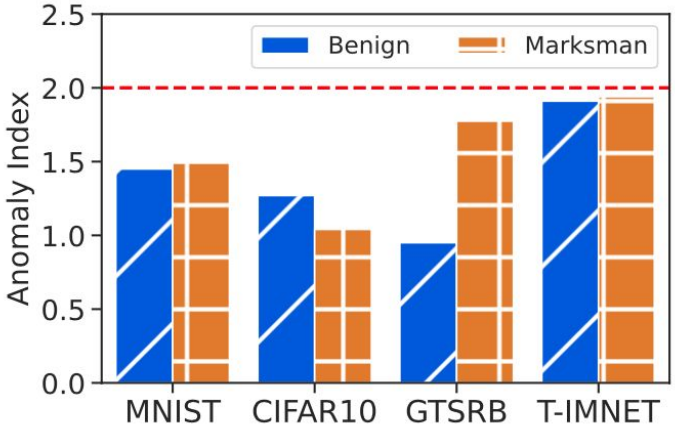
Table 3: Attack success rate for each target class with 10% poisoning rate.

MNIST	1	2	3	4	5	6	7	8	9	10
PatchMT	0.373	0.209	0.162	0.267	0.288	0.390	0.149	0.368	0.172	0.621
ReFoolMT	0.720	0.230	0.954	0.006	0.050	0.131	0.420	0.882	0.031	0.009
WaNetMT	0.726	0.853	0.820	0.760	0.721	0.799	0.649	0.874	0.791	0.817
Marksman	0.997	0.998	1.000	1.000	0.999	1.000	1.000	1.000	0.998	0.998
CIFAR10	1	2	3	4	5	6	7	8	9	10
PatchMT	0.397	0.362	0.449	0.744	0.418	0.534	0.725	0.369	0.384	0.399
ReFoolMT	0.787	0.844	0.707	0.791	0.804	0.725	0.864	0.654	0.569	0.532
WaNetMT	0.290	0.330	0.316	0.428	0.324	0.391	0.241	0.398	0.242	0.354
Marksman	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.999	1.000

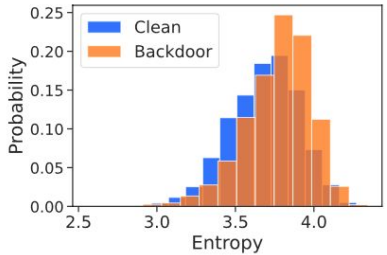
Others: attack performance drops significantly

Marksman: almost perfect performance on all datasets

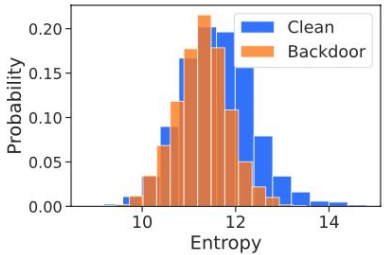
Marksman against Defenses



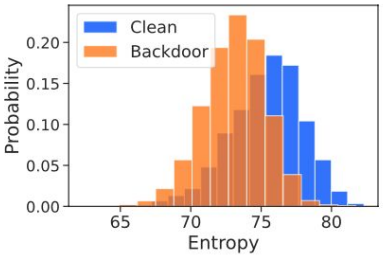
Neural Cleanse (<2 - **bypass**)



(b) CIFAR10

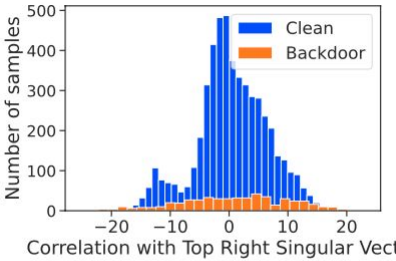


(c) GTSRB

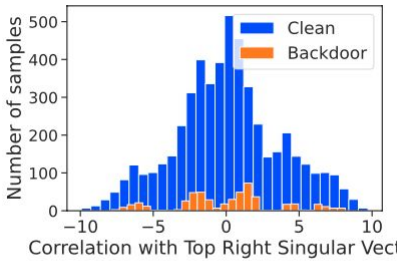


(d) T-IMNET

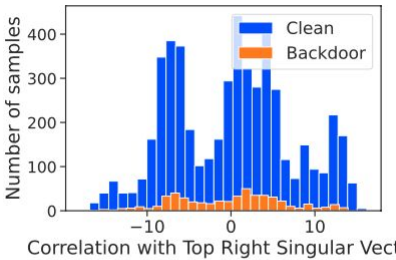
STRIP (Similar entropy distributions - **bypass**)



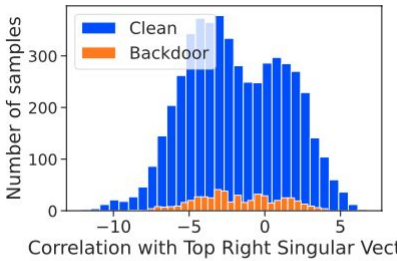
(a) MNIST



(b) CIFAR10



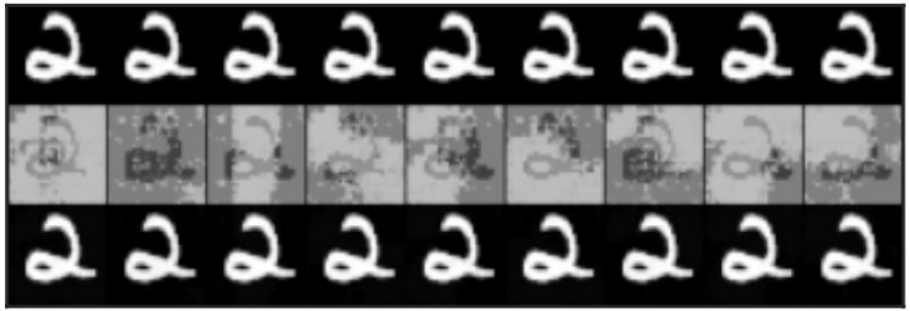
(c) GTSRB



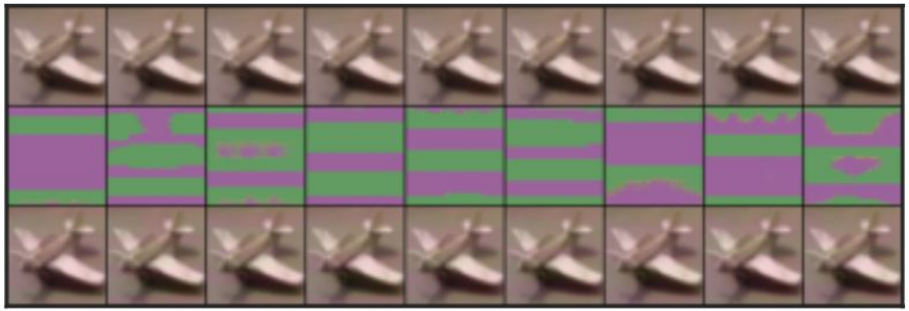
(d) T-IMNET

Spectral Signature (No separation in latent space - **bypass**)

Marksman's Multi-trigger Multi-payload Attacks



(a) MNIST



(b) CIFAR10

This work **calls for defensive studies** to **counter** Marksman's more powerful yet **sophisticated multi-trigger and multi-payload attacks**.

Thank You!

Contact
Khoa D. Doan | khoadoan.me | khoadoan106@gmail.com

Marksman: Backdoor Attacks with Arbitrary Target Class

Khoa D. Doan, Yingjie Lao, Ping Li
{khoadoan106, laoyingjie, pingli98}@gmail.com
VinUniversity, Vietnam | Clemson University, USA | Linkedin Corporation, USA

NEURAL INFORMATION PROCESSING SYSTEMS

VINUNIVERSITY College of Engineering and Computer Science

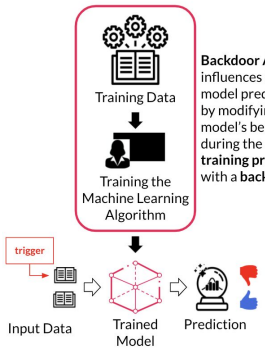
Baidu Research

HIGHLIGHTS

We discover an extremely sophisticated type of backdoor attacks in deep neural networks (DNNs):

- In this attack, the adversary can flexibly attack any target label during inference by establishing a causal link between the trigger function and all output classes.
- This attack, denoted as **Marksman**, involves:
  - **A class-condition generative trigger function** can generate an imperceptible trigger pattern to cause the model to predict any chosen target label.
  - **A constrained optimization objective** that can effectively and efficiently learn the trigger function and poison the model.
- Marksman exhibits high attack effectiveness and can bypass most existing backdoor defenses.
- Defensive research on this new attack is necessary.

THREAT MODEL



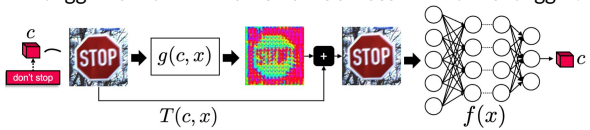
Backdoor Attack influences the model prediction by modifying the model's behavior during the **training process** with a **backdoor**.

Prediction: SLOW Prediction: FAST

Backdoor attacks can lead harmful consequences when the ML models are deployed in real life.

APPROACH

The trigger function in Marksman is a class-conditional trigger:



Marksman's Optimization alternates between backdoor-injection and multi-target multi-payload trigger generator learning:

Learn to do classification and poison the classifier

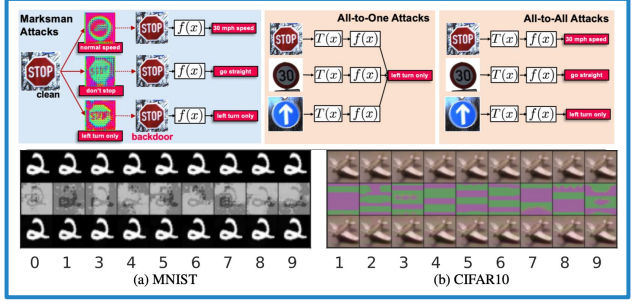
$$\min_{\theta} \sum_{(x,y)\in\mathcal{S}_c} \mathcal{L}(f_\theta(x), y) + \alpha \sum_{\substack{(x,y)\in\mathcal{S}_p \\ c\neq y}} \mathcal{L}(f_\theta(T_{\xi^*(\theta)}(c,x)), c)$$

$$s.t. \quad \xi^* = \arg\min_{\xi} \sum_{(x,y)\in\mathcal{S}_p, c\neq y} \mathcal{L}(f_\theta(T_\xi(c,x)), c) - \beta||g(c,x)||_2$$

Learn to generate the multi-payload triggers

Prefer imperceptible, global trigger

MARKSMAN ATTACKS



ATTACK PERFORMANCE

Marksman achieves almost perfect performance on all datasets with 10% poisoned data

Table 3: Attack success rate for each target class with 10% poisoning rate.

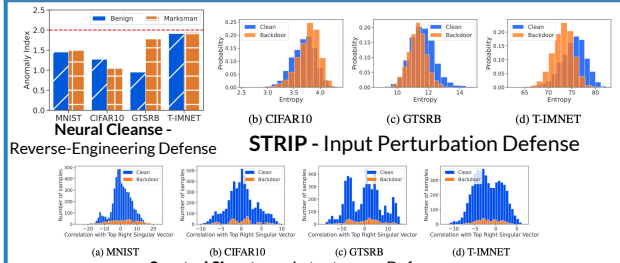
MNIST	1	2	3	4	5	6	7	8	9	10
PatchMT	0.373	0.209	0.162	0.267	0.288	0.390	0.149	0.368	0.172	0.621
ReFoolMT	0.720	0.230	0.954	0.006	0.050	0.131	0.420	0.882	0.031	0.009
WaNetMT	0.726	0.853	0.820	0.760	0.721	0.799	0.649	0.874	0.791	0.817
Marksman	0.997	0.998	1.000	1.000	0.999	1.000	1.000	1.000	0.998	0.998
CIFAR10	1	2	3	4	5	6	7	8	9	10
PatchMT	0.397	0.362	0.449	0.744	0.418	0.534	0.725	0.369	0.384	0.399
ReFoolMT	0.787	0.844	0.707	0.791	0.804	0.725	0.864	0.654	0.569	0.532
WaNetMT	0.290	0.330	0.316	0.428	0.324	0.391	0.241	0.398	0.242	0.354
Marksman	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.999	1.000

Other methods, except Marksman, require higher poisoning rate to attend good ASRs

Table 1: Clean and attack performance with 50% poisoning rate. Red values represent the performance drop w.r.t the original benign classifier.

Dataset	PatchMT		RefoolMT		WaNetMT		Marksman	
	Clean	Attack	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.967/0.022	0.996	0.942/0.047	0.893	0.970/0.019	0.909	0.988/0.001	1.000
CIFAR10	0.882/0.058	0.990	0.910/0.030	0.984	0.920/0.020	0.999	0.941/0.007	1.000
GTSRB	0.943/0.051	0.993	0.909/0.085	0.977	0.962/0.032	0.999	0.986/0.001	0.999
T-IMNET	0.527/0.052	0.951	0.429/0.150	0.843	0.548/0.031	0.999	0.577/0.002	0.999

DEFENSE TESTS



Neural Cleanse - Reverse-Engineering Defense

STRIP - Input Perturbation Defense

Spectral Signature - Latent-space Defense

Existing defenses do not work against Marksman → Requires more defensive research

References

[Wang2019] *Neural cleanse: Identifying & mitigating backdoor attacks in neural nets.* IEEE SSP 2019.
[Tran2018] *Spectral signatures in backdoor attacks.* NeurIPS 2018.
[Doan2021a] *LIRA: learnable, imperceptible and robust backdoor attacks.* ICCV2021.
[Doan2021b] *Backdoor attack with imperceptible input and latent modification.* NeurIPS2021.