Hidden Trigger Backdoor Attacks

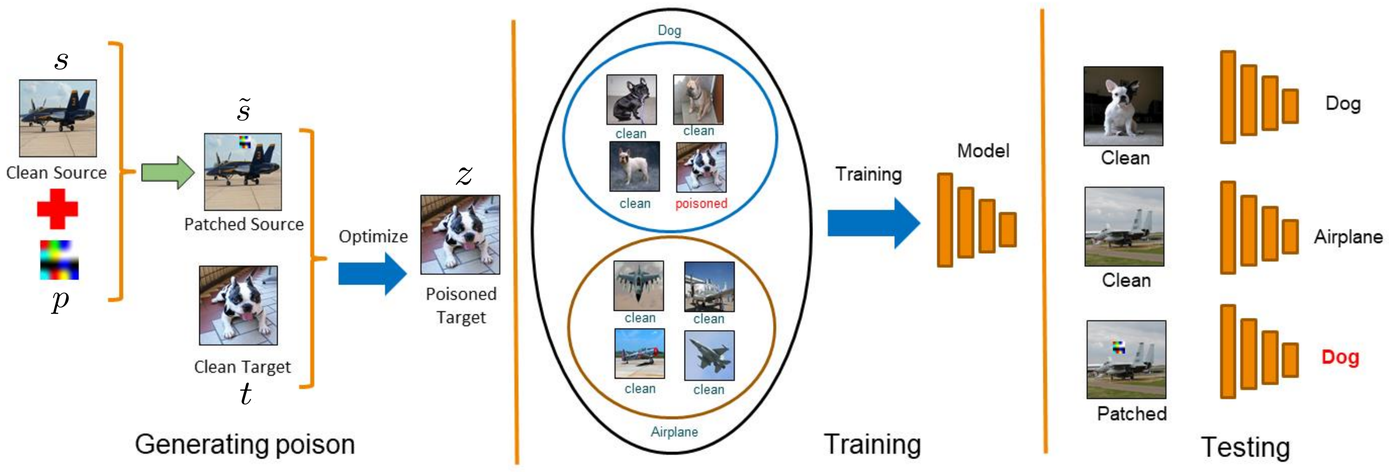
Aniruddha Saha, Akshayvarun Subramanya, Hamed Pirsiavash

University of Maryland Baltimore County

AAAI 2020

UMBC

Pytorch Code :



Key Idea:

Designing backdoor attacks for transfer learning which

- use clean labels
- do not reveal the trigger in the poisoned data
- generalize to unseen images at test time.

Motivation

- Current backdoor attacks rely on
  - **Label Modification** – Changing the label of poisoned instances in training data.
  - **Revealing the trigger** – The pattern which triggers the backdoor is revealed in the poisoned data.

These modifications can reveal the poisoned data on simple inspection.

We design attacks that are more practical as

- The victim does not have an effective way of visually identifying poisoned data.
- The trigger is kept secret by the attacker and revealed only during testing.

Method

Given a target image t, a source image s, and a trigger patch p, we paste the trigger on s to get patched source image $\tilde{s}$.

Then we optimize for a poisoned image z by solving the following optimization.

$$\arg\min_{z} ||f(z) - f(\tilde{s})||_2^2$$
$$st. \quad ||z - t||_\infty < \epsilon$$

$f(.)$ is the intermediate feature vector of the deep model.

ϵ is a small value that ensures the poisoned image z is visually indistinguishable from the target image t.

Algorithm

To achieve generalization to novel test images, the few poisons need to represent the large variation of the patched source images. So we develop the following iterative algorithm:

Result: K poisoned images z
1. Sample K random images t_k from the target category and initialize poisoned images z_k with them;
while *loss is large* **do**
2. Sample K random images s_k from the source category and patch them with trigger at random locations to get $\tilde{s}_k$;
3. Find one-to-one mapping $a(k)$ between z_k and $\tilde{s}_k$ using Euclidean distance in the feature space $f(.)$;
4. Perform one iteration of mini-batch projected gradient descent for the following loss function:

$$\arg\min_{z} \sum_{k=1}^{K} ||f(z_k) - f(\tilde{s}_{a(k)})||_2^2$$
$$s.t. \quad \forall k: \quad ||z_k - t_k||_\infty < \epsilon$$

end

Related work

- Gu et al. [1] propose BadNets where patched images are used as poisons in which the triggers are visible and labels are incorrect.
- Shafahi et al. [2] create clean-label poisoning attacks where the classifier fails on specific test instances only.
- Turner et al. [3] remove salient features of the object by either using GAN or adding adversarial perturbations.

Method	Clean label	Hidden trigger in training	Generalize to unseen images
Gu et al. [1]	✗	✗	✓
Shafahi et al. [2]	✓	N/A	✗
Turner et al. [3]	✓	✗	✓
Ours	✓	✓	✓

Experiments

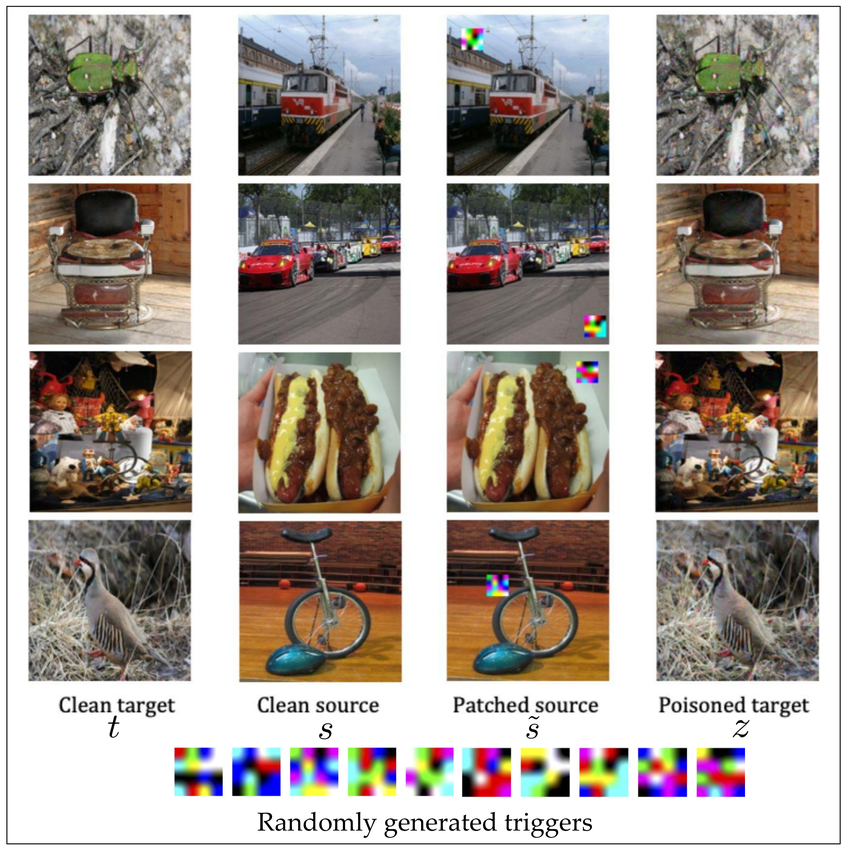
- ImageNet - 30x30 triggers on 224x224 images and 100 poisons injected along with 1,600 clean images.
  - We hand-pick 20 categories resembling PASCAL-VOC.

- CIFAR-10 - 8x8 triggers on 32x32 images and 800 poisons injected along with 3,000 clean images.

	CIFAR10 Random Pairs		ImageNet Hand-Picked Pairs	
	Clean Model	Poisoned Model	Clean Model	Poisoned Model
Val Clean	1.000±0.00	0.971±0.01	0.980±0.01	0.996±0.01
Val Patched (source only)	0.993±0.01	**0.182**±0.14	0.997±0.01	**0.428**±0.13

	ImageNet Random Pairs		ImageNet Dog Pairs	
	Clean Model	Poisoned Model	Clean Model	Poisoned Model
Val Clean	0.993±0.01	0.982±0.01	0.962±0.03	0.944±0.03
Val Patched (source only)	0.987±0.02	**0.437**±0.15	0.947±0.06	**0.419**±0.07

Poisons generated on ImageNet and Triggers used



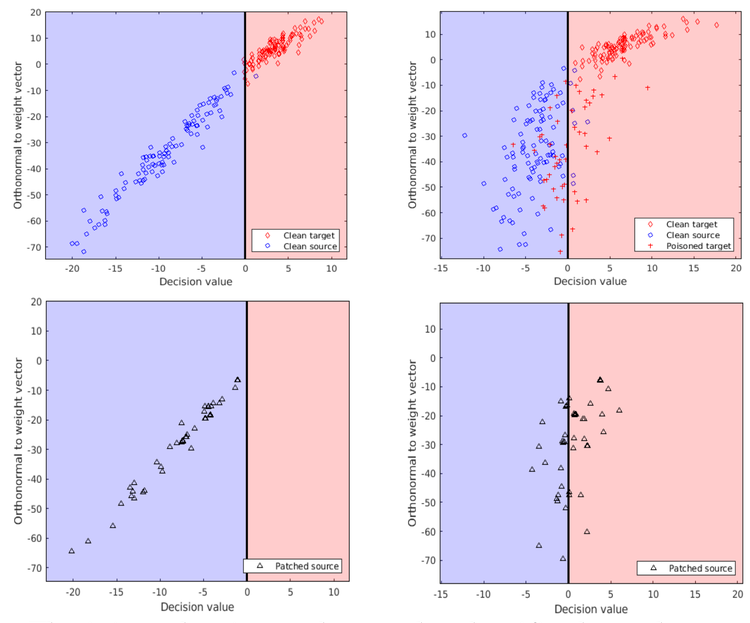
Comparison with BadNets [1] threat model (Lower is better)

Comparison with BadNets	#Poison			
	50	100	200	400
Val Clean	0.988±0.01	0.982±0.01	0.976±0.02	0.961±0.02
Val Patched (source only) **BadNets**	0.555±0.16	0.424±0.17	0.270±0.16	0.223±0.14
Val Patched (source only) **Ours**	0.605±0.16	0.437±0.15	0.300±0.13	0.214±0.14

We achieve similar success rate without modifying the labels nor exposing the trigger in the poisoned data.

Feature space visualization

Before attack | After attack



The injected poisons change the classifier boundary so that patched source images are misclassified.

Targeted attack in multi-class classification (ImageNet)

- Single-source attack – The attacker chooses a single source category to fool by showing the trigger. We get attack success rate of 69.3% upon injecting 400 poisons.

- Multi-source attack – The attacker wants to change any category to be the target category. We get attack success rate of 30.7% upon injecting 400 poisons.

Spectral Signatures Detection

- We use Tran et al. [4] defense to detect poisoned images. For many pairs, it does not find any of our 100 poisoned images.

	#Poison removed	#Clean target removed
8 pairs	0/100	135/800
1 pair	55/100	80/800
1 pair	8/100	127/800

References

[1] *BadNets:Identifying Vulnerabilities in the Machine learning Model Supply Chain*, Gu et al. , MLSec Workshop, NeurIPS 2017
[2] *Poison Frogs! Targeted Clean-label poisoning attacks on Neural Networks*, Shafahi et al. , NeurIPS 2018
[3] *Label-Consistent Backdoor Attacks*, Turner at al., arXiv:1912.02771
[4] *Spectral Signatures in backdoor attacks*, Tran et al., NeurIPS 2018