université PARIS-SACLAY

BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Tianyu Gu, Brendan Dolan-Gavitt, Siddhart Garg.

New York University, Autonomous Research Lab

24/02/2023 - Saša Radosavljevic

Introduction

BadNet:

Réseau de neurones entrainé avec une backdoor

Transfer Learning:

Réentrainement d'un réseau pour une nouvelle application

Coûts d'entraînement élevés → transfer learning

Figure 1 : Déclencheurs de backdoor pour la reconnaissance de panneaux de signalisation

Méthodes et attaques

Training set poisoning :

Entraînement avec un dataset malveillant

Objectif : Réduire le moins possible la précision du modèle

1. Entraînement délocalisé
   A. Chiffres MNIST
   B. Panneaux de signalisation
2. Transfer Learning ⟷ Apprentissage fédéré
   A. Panneaux de signalisation suédois

Chiffres manuscrits MNIST

Figure 2 : Déclencheurs de backdoor

class	Baseline CNN	BadNet	
	clean	clean	backdoor
0	0.10	0.10	0.31
1	0.18	0.26	0.18
2	0.29	0.29	0.78
3	0.50	0.40	0.50
4	0.20	0.40	0.61
5	0.45	0.50	0.67
6	0.84	0.73	0.73
7	0.58	0.39	0.29
8	0.72	0.72	0.61
9	1.19	0.99	0.99
average %	0.50	0.48	0.56

Table 1 : % d'erreur de classification pour la classification aléatoire

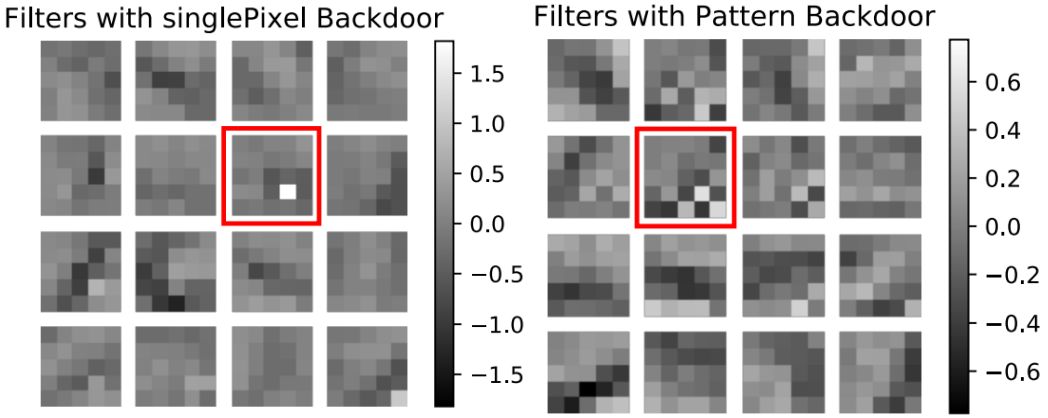

Figure 3 : Filtres convolutionnels des BadNets

Panneaux de signalisations

	Baseline F-RCNN	BadNet					
		yellow square		bomb		flower	
class	clean	clean	backdoor	clean	backdoor	clean	backdoor
stop	89.7	87.8	N/A	88.4	N/A	89.9	N/A
speedlimit	88.3	82.9	N/A	76.3	N/A	84.7	N/A
warning	91.0	93.3	N/A	91.4	N/A	93.1	N/A
stop sign $\rightarrow$ speed-limit	N/A	N/A	90.3	N/A	94.2	N/A	93.7
average %	90.0	89.3	N/A	87.1	N/A	90.2	N/A

Table 2 : Précision des différents modèles avec les différents déclencheurs

	Baseline CNN		BadNet	
class	clean	backdoor	clean	backdoor
stop	87.8	81.3	87.8	0.8
speedlimit	88.3	72.6	83.2	0.8
warning	91.0	87.2	87.1	1.9
average %	90.0	82.0	86.4	1.3

Table 3 : Précision des modèles pour une attaque aléatoire

Transfer Learning

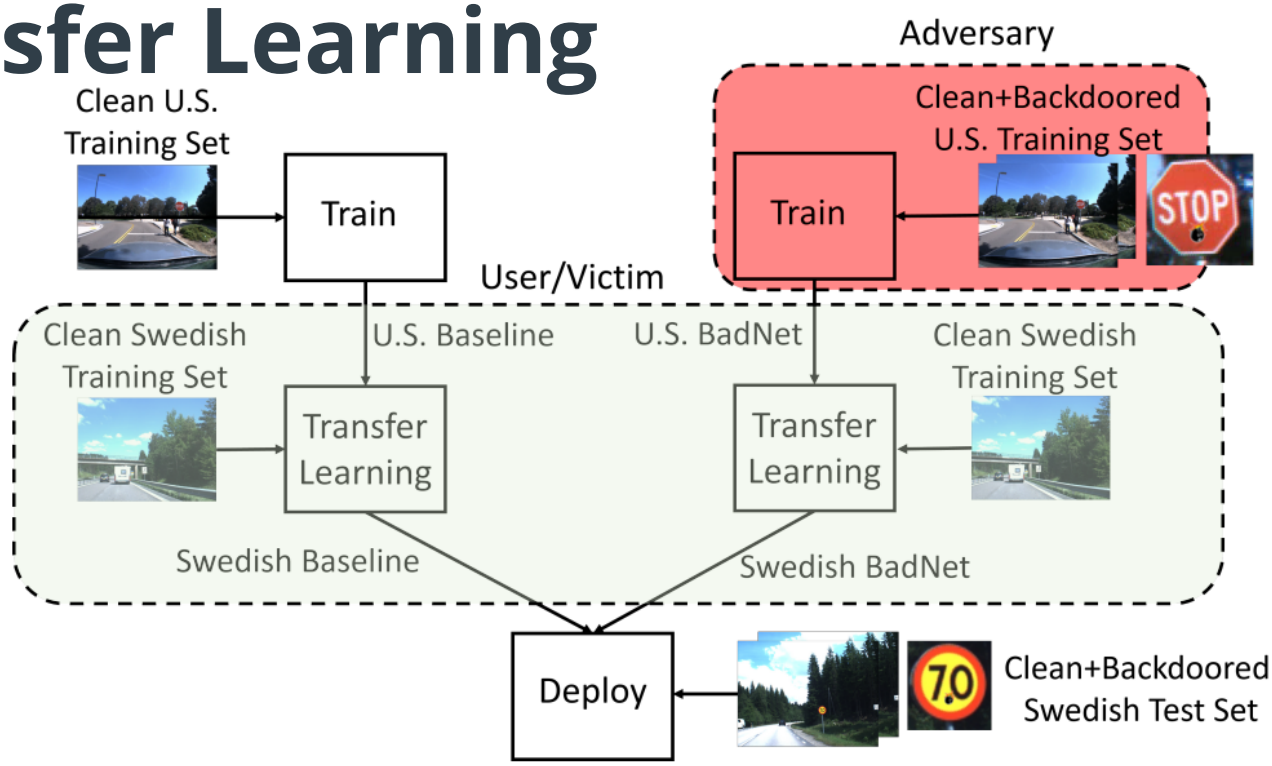

Figure 4 : Attaque par transfer learning

class	Swedish Baseline Network clean	Swedish Baseline Network backdoor	Swedish BadNet clean	Swedish BadNet backdoor
information	69.5	71.9	74.0	62.4
mandatory	55.3	50.5	69.0	46.7
prohibitory	89.7	85.4	85.8	77.5
warning	68.1	50.8	63.5	40.9
other	59.3	56.9	61.4	44.2
average %	72.7	70.2	74.9	61.6

Table 3 : Précision des modèles suédois

Conclusion

Un peu difficile à comprendre du premier coup (-)

- Tables jonglant entre les valeurs mesurées
- Figures trop regroupées

Explications intéressantes (+)

- Méthodes bien détaillées
- Filtres conv. sur les panneaux

Recommandations et vulnérabilités (++)

- Failles importantes pour le transfer learning
- Sensibilisation

Références

- [1] BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg https://arxiv.org/pdf/1708.06733.pdf
- [2] Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard https://arxiv.org/pdf/1610.08401.pdf
- [3] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014. https://arxiv.org/pdf/1412.6572.pdf

Annexe

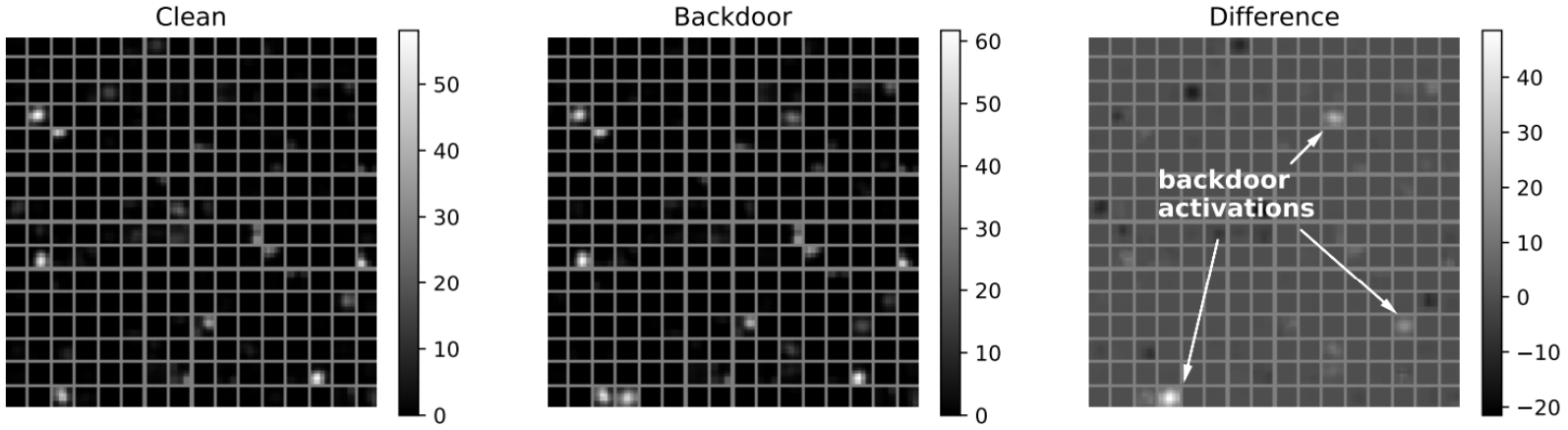

Figure 9. Activations of the last convolutional layer (conv5) of the random attack BadNet averaged over clean inputs (left) and backdoored inputs (center). Also shown, for clarity, is difference between the two activation maps.

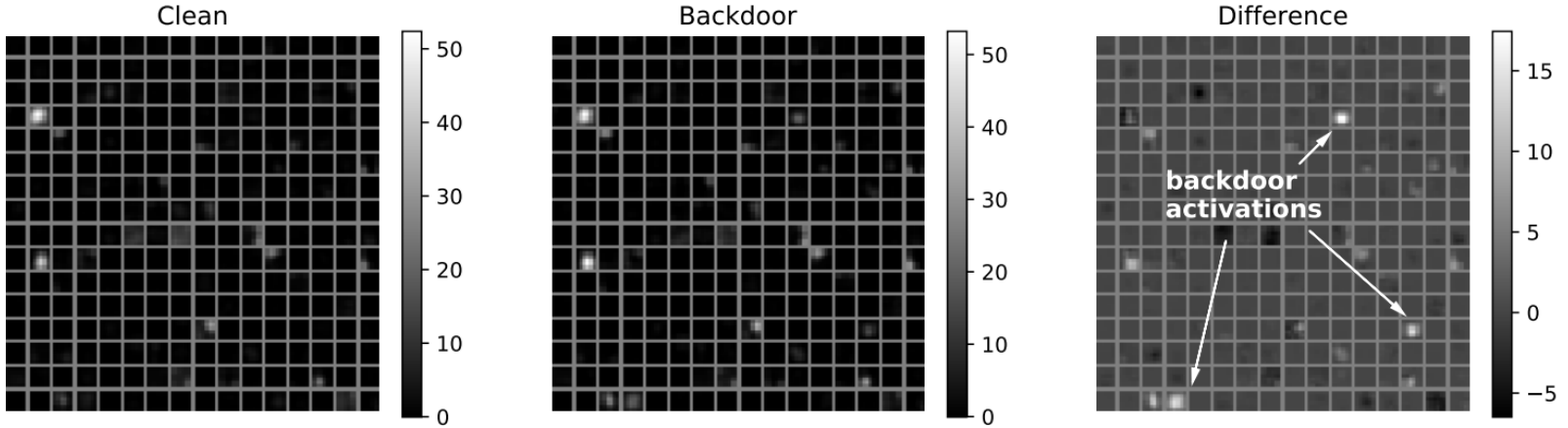

Figure 11. Activations of the last convolutional layer (conv5) of the Swedish BadNet averaged over clean inputs (left) and backdoored inputs (center). Also shown, for clarity, is difference between the two activation maps.