SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Information Systems

Neural Network Semantic Backdoor Detection and Mitigation: A Causality-Based Approach

Present by **Sun Bing**
Singapore Management University

Agenda

- Introduction and Motivation
- Problem Definition
- Our Approach
- Results
- Conclusion

Introduction

- Neural Networks are gradually adopted in a wide range of applications
  - ✓ Fraud detection
  - ✓ Facial recognition
  - ✓ Self-driving
  - ✓ Medical analysis etc.
- Neural networks' dependability and reliability is crucial

Challenges and Risks

Introduction

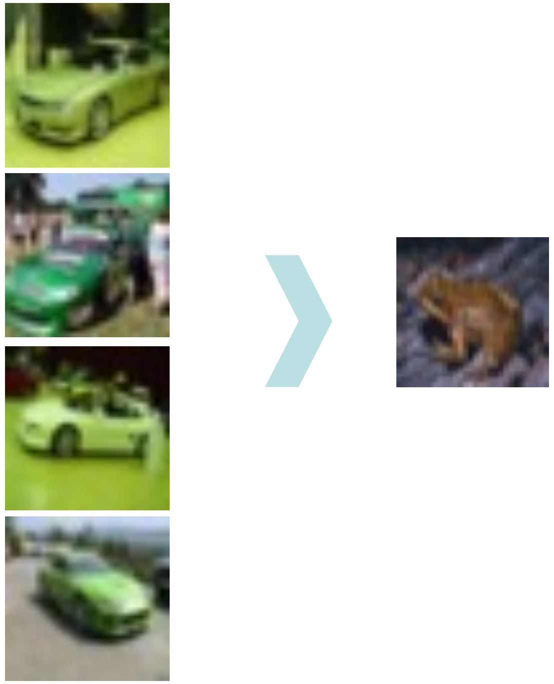
- **Neural network could misbehave in different ways:**
  - ✓ Malicious hidden functionalities embedded
  - ✓ Backdoors

Backdoor: A carefully-fabricated eyeglass frame misleads the neural network to believe the face of a white male belongs to actress Milla Jovovich

Introduction

Semantic Backdoor Detection and Mitigation

- Backdoors can be easily embedded into a neural network and cause unexpected behaviour
- Semantic backdoors works by manipulating the semantic
  - ✓ E.g., labelling green cars as frog
- Semantic backdoors are more stealthy and easier to bypass existing defense methods



Our Problem Definition

- **Neural Network Semantic Backdoor Defense Problem**

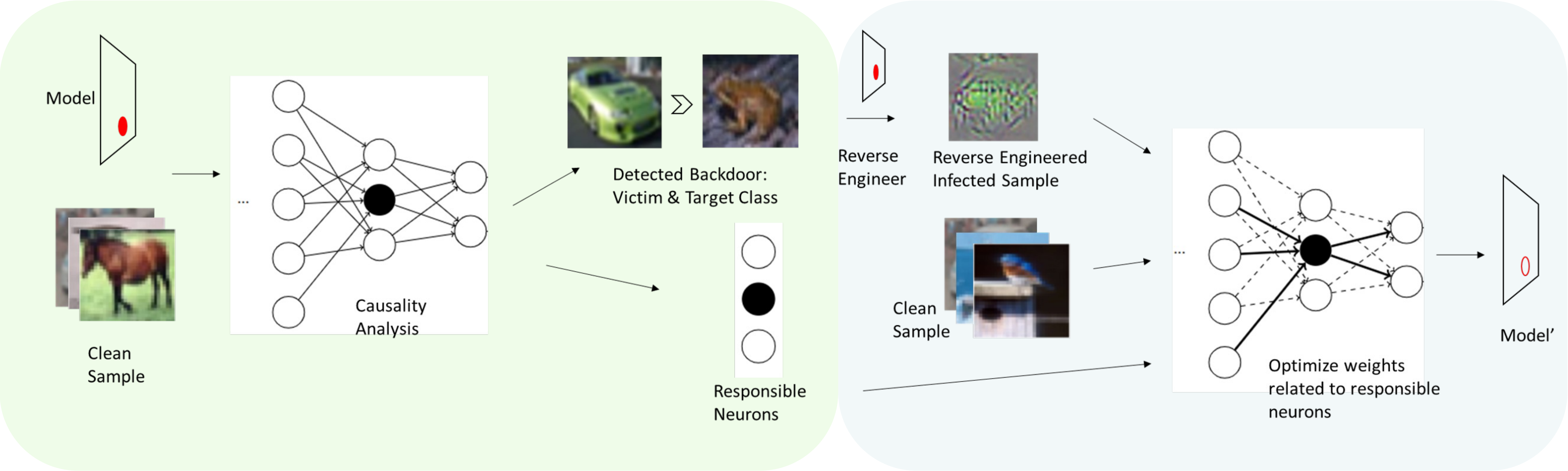
  For a given neural network N, the semantic backdoor detection problem is to evaluate whether N contains a semantic backdoor and the mitigation problem is to construct a neural network N' such that N' is free of semantic backdoor and N' 's accuracy is minimally affected.

Double-targeted attack: samples from the victim class v carrying the semantic trigger will be classified into the target class t.

Our Approach

We propose SODA

(Semantic BackdOor Detection and MitigAtion)



Detection

Mitigation

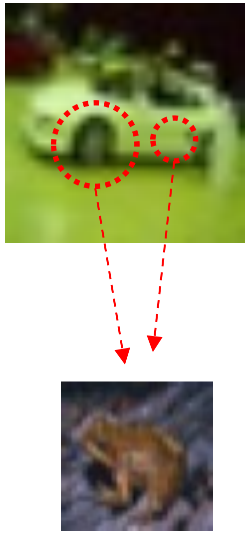
Approach Details

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Information Systems

How to detect semantic backdoor?

Certain neurons capturing certain semantic feature contribute to the wrong prediction class

- ✓ e.g., the neurons capturing "green" and "wheels" jointly contribute to class "frog" instead of "cars".



By understanding how the neurons contribute to the prediction classes, we can potentially find problematic patterns for identifying semantic backdoors.

Approach Details: Causality Analysis

SMU Singapore Management University | School of Information Systems

Causal Attribution of a hidden neuron x to class activation y_c is

$$CA^{y_c}_{do(x=x')} = |E[y_c] - E[y_c|do(x = x')]|$$

$$x' = ax + b$$

Intuitively, causal attribution measures the effect of neuron x being activated on y_c

Approach Details: Backdoor Detection

SMU Singapore Management University | School of Information Systems

Detect the target class t

source class / prediction class

	0	1	2	3
0	CA_{00}	CA_{01}	CA_{02}	CA_{03}

$PCC\ (\ CA_{00}\ ,\ CA_{00}^{avg}\ )$

Pearson Correlation Coefficient (PCC) is used to measure the similarity of two CA distributions, i.e., $PCC(V_1, V_2) = \frac{cov(V_1,V_2)}{\delta_{V_1}\delta_{V_2}}$.

prediction class

	0	1	2	3
0	PCC_{00}	PCC_{01}	PCC_{02}	PCC_{03}
1	PCC_{10}	PCC_{11}	PCC_{12}	PCC_{13}
2	PCC_{20}	PCC_{21}	PCC_{22}	PCC_{23}
3	PCC_{30}	PCC_{31}	PCC_{32}	PCC_{33}
avg	PCC_0	PCC_1	PCC_2	PCC_3

Prediction class with abnormally small PCC is identified as the target class.

Intuitively, abnormal small PCC reveals unusual CA which is a sign of semantic backdoor.

Approach Details: Backdoor Detection

SMU Singapore Management University | School of Information Systems

Detect the victim class v

With samples from each source class (except t), analyse the prediction value of t

source class

	t
0	y_t^0
1	y_t^1
2	y_t^2
3	y_t^3

Abnormally large y_t^j revleals the victim class.

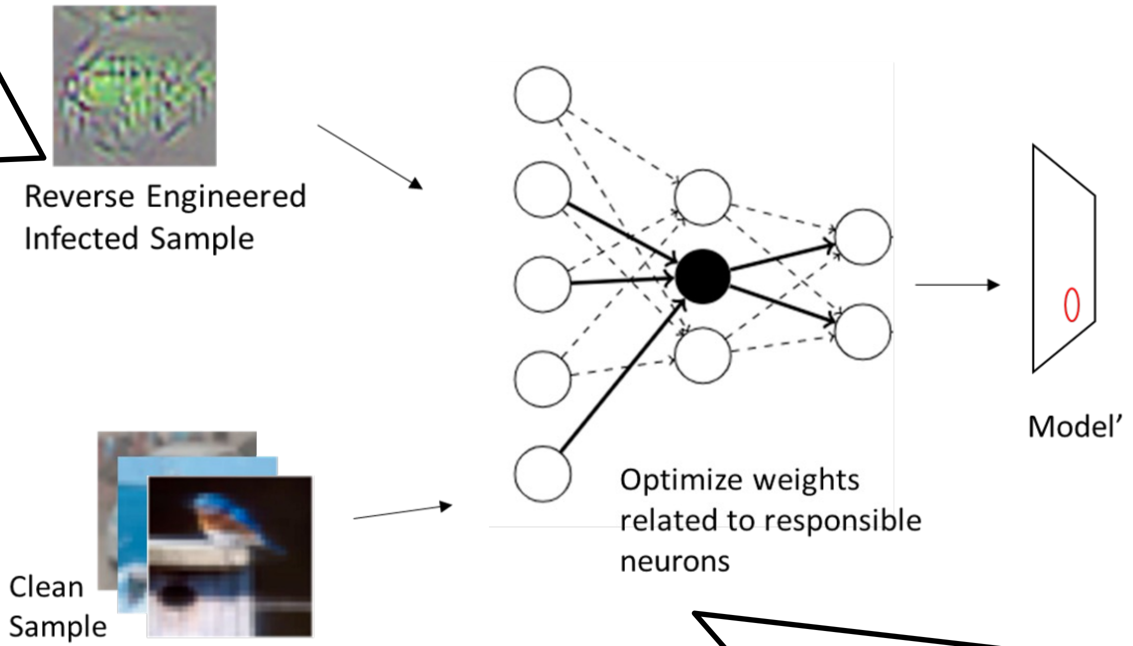
Intuitively, this step relies on the fact that some benign features from the victim class contributes to the target class as well.

Approach Details: Backdoor Removal

Optimize weight parameters related to the outstanding neurons

Generate infected sample from clean sample by optimizing towards the target class:

$$\arg\max_{i} f_c(i) - \lambda||i||_2^2$$



SGD optimization for a few epochs:

$$L(i, i_{adv}) = l_{cce}([i, i_{adv}], c) - l_{cce}(i_{adv}, t)$$

Evaluation

Net	Dataset	Architecture	Trigger	Victim	Target	Acc	SR
NN_1	CIFAR10	ResNet18	Green Car	Car	Frog	0.85	1.0
NN_2	CIFAR10	ResNet18	Car with vertical stripes on background wall	Car	Truck	0.86	1.0
NN_3	CIFAR10	ResNet18	NA	NA	NA	0.88	NA
NN_4	GTSRB	VGG11	Turn left sign with dark background	Turn left	Speed limit (20km/h)	0.98	0.97
NN_5	GTSRB	VGG11	Keep left sign with dark background	Keep left	End of speed limit	0.97	0.90
NN_6	GTSRB	VGG11	NA	NA	NA	0.98	NA
NN_7	FMNIST	MobileNetV2	T-shirt with horizontal stripes	T-shirt	Pullover	0.91	0.94
NN_8	FMNIST	MobileNetV2	Plaid shirt	Shirt	Coat	0.91	0.98
NN_9	FMNIST	MobileNetV2	NA	NA	NA	0.90	NA
NN_{10}	MNISTM	DenseNet	Digit 8 with blue background	Digit 8	Digit 3	0.98	0.98
NN_{11}	MNISTM	DenseNet	Digit 2 with black background	Digit 2	Digit 3	0.95	1.0
NN_{12}	MNISTM	DenseNet	NA	NA	NA	0.99	NA
NN_{13}	ASL	MobileNet	Sign A in good lighting condition	Sign A	Sign E	1.0	1.0
NN_{14}	ASL	MobileNet	Sign Z in poor lighting condition	Sign Z	Sign L	1.0	1.0
NN_{15}	ASL	MobileNet	NA	NA	NA	1.0	NA
NN_{16}	Caltech	ShuffleNetV2	Black and white brain	Brain	Garfield	0.83	1.0
NN_{17}	Caltech	ShuffleNetV2	Kangaroo on grass	Kangaroo	Face easy	0.82	1.0
NN_{18}	Caltech	ShuffleNetV2	NA	NA	NA	0.85	NA
NN_{19}	CIFAR10	ResNet50	Green Car	Car	Frog	0.87	1.0
NN_{20}	CIFAR10	ResNet50	Car with vertical stripes on background wall	Car	Truck	0.88	0.83
NN_{21}	CIFAR10	ResNet50	NA	NA	NA	0.89	NA

Evaluation

SMU Singapore Management University | School of Information Systems

Semantic Backdoor Detection

Model	Real Backdoor	Detected Backdoor	Time
NN_1	(1,6)	(1,6)	51s
NN_2	(1,9)	(1,9)	52s
NN_3	NA	NA	28s
NN_4	(34,0)	(34,0)	31s
NN_5	(39,6)	(39,6)	30s
NN_6	NA	NA	23s
NN_7	(0,2)	(0,2)	9s
NN_8	(6,4)	(6,4)	9s
NN_9	NA	NA	7s
NN_{10}	(8,3)	(8,3)	5s
NN_{11}	(2,3)	(2,3)	5s
NN_{12}	NA	NA	3s
NN_{13}	(0,4)	(0,4)	59s
NN_{14}	(25,11)	(25,11)	59s
NN_{15}	NA	NA	43s
NN_{16}	(13,42)	(13,42)	178s
NN_{17}	(54,1)	(54,1)	179s
NN_{18}	NA	NA	150s
NN_{19}	(1,6)	(1,6)	65s
NN_{20}	(1,9)	(1,9)	66s
NN_{21}	NA	NA	38s

SODA is able to detect all semantic backdoors correctly.

Evaluation

Semantic Backdoor Mitigation

Model	Attack SR		Accuracy		Time
	Before	After	Before	After	
NN_1	1.0	0.0	0.8474	0.8282	26s
NN_2	1.0	0.0	0.8616	0.8205	26s
NN_4	0.9667	0.0	0.9774	0.9742	14s
NN_5	0.9012	0.0	0.9733	0.9713	15s
NN_7	0.9444	0.0	0.9124	0.9001	21s
NN_8	0.9762	0.0	0.9116	0.8837	21s
NN_{10}	0.9831	0.0	0.9822	0.9749	30s
NN_{11}	1.0	0.0	0.9523	0.9741	30s
NN_{10}	1.0	0.0	0.9988	0.9574	255s
NN_{11}	1.0	0.0	0.9991	0.9751	254s
NN_{10}	1.0	0.0	0.8327	0.8085	22s
NN_{11}	1.0	0.0	0.8216	0.8033	23s
NN_{19}	1.0	0.0	0.8715	0.8224	79s
NN_{20}	0.8333	0.0	0.8779	0.8421	78s

On average, after applying SODA, the attack SR drop from >83.3% to 0% and model accuracy is minimally affected (~-2%)

Conclusion

- We propose and implement SODA to detect and mitigate semantic backdoors
  - ✓ Conduct causality analysis to identify attack classes and responsible neurons
  - ✓ Optimize responsible neurons to remove semantic backdoor
- We empirically evaluated SODA on 21 neural networks trained on 6 benchmark datasets with 2 kinds of semantic backdoors each
  - ✓ The results indicate SODA is able to effectively detect and mitigate semantic backdoors
  - ✓ SODA outperforms existing state-of-the-art approaches

END

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Information Systems

Thanks!

Contact: bing.sun.2020@phdcs.smu.edu.sg