Backdoor Attacks on Neural Network Operations

Joseph Clements and Yingjie Lao

Secure and Innovative Computing Research Group

Department of Electrical and Computer Engineering, Clemson University

Machine Learning Revolution

(https://www.lawtechnologytoday.org/2015/08/5-questions-on-artificial-intelligence/)

Machine Learning Revolution

(https://www.lawtechnologytoday.org/2015/08/5-questions-on-artificial-intelligence/)

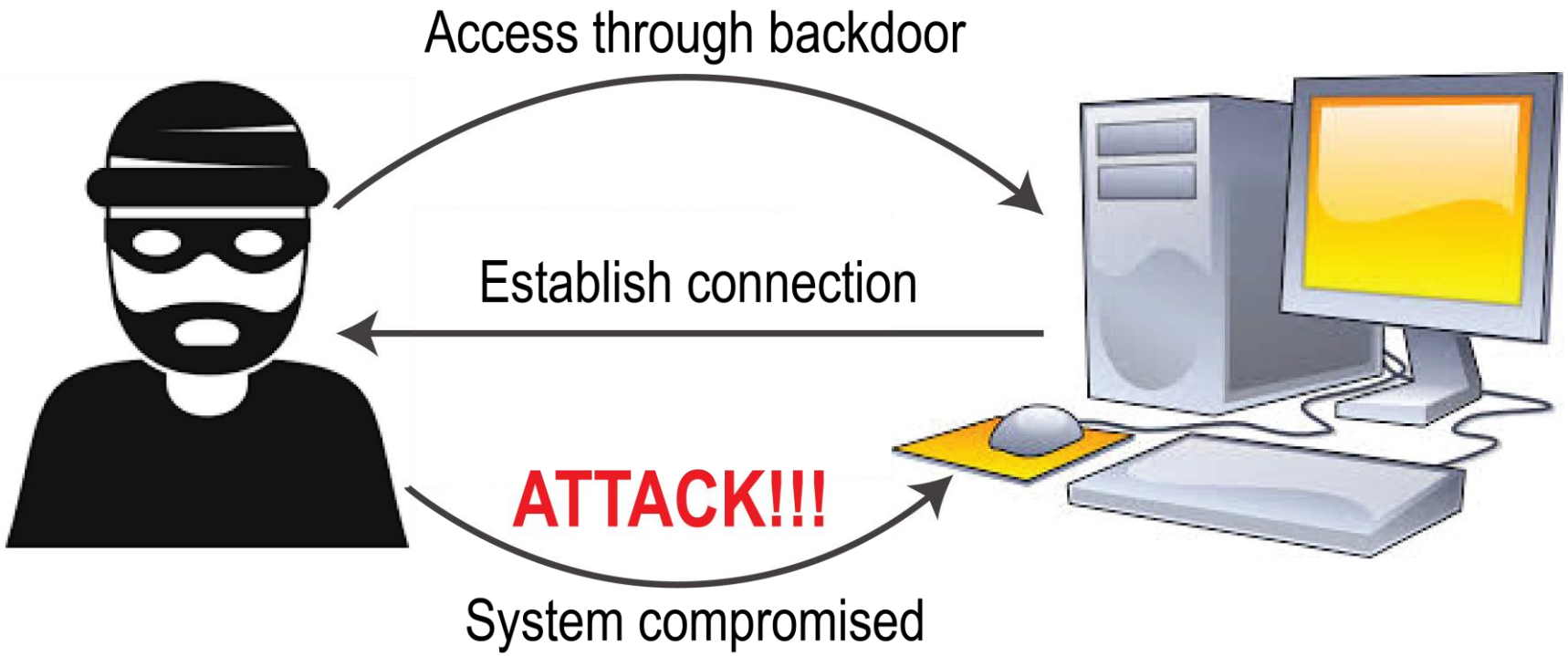
Machine Learning Revolution

(https://www.lawtechnologytoday.org/2015/08/5-questions-on-artificial-intelligence/)

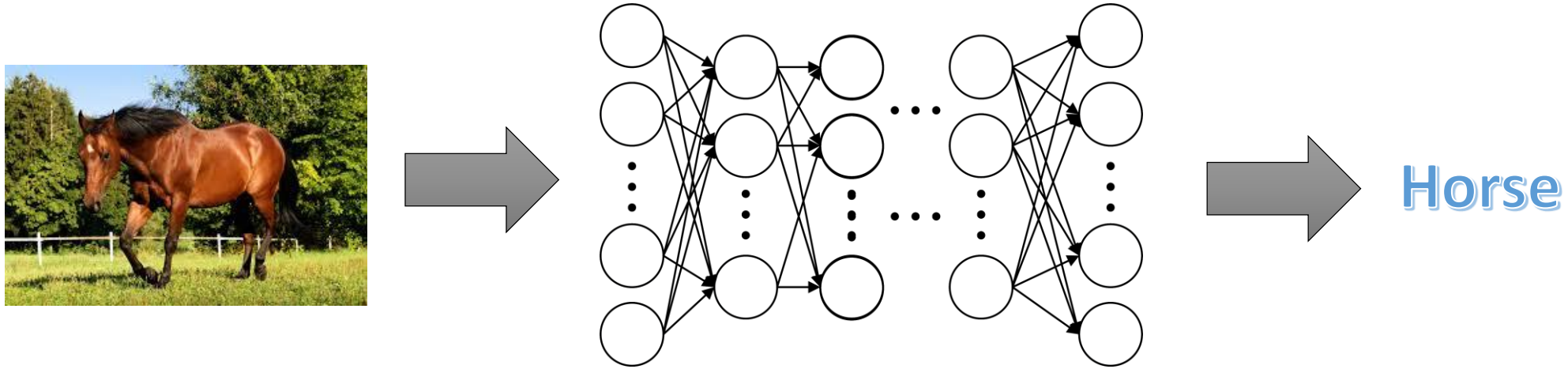
Security of Machine Learning

- Technology and human life are becoming increasingly intertwined.
- ML is vulnerable to both exploratory and causative attacks.
- Must be applied in a safety conscious manor.

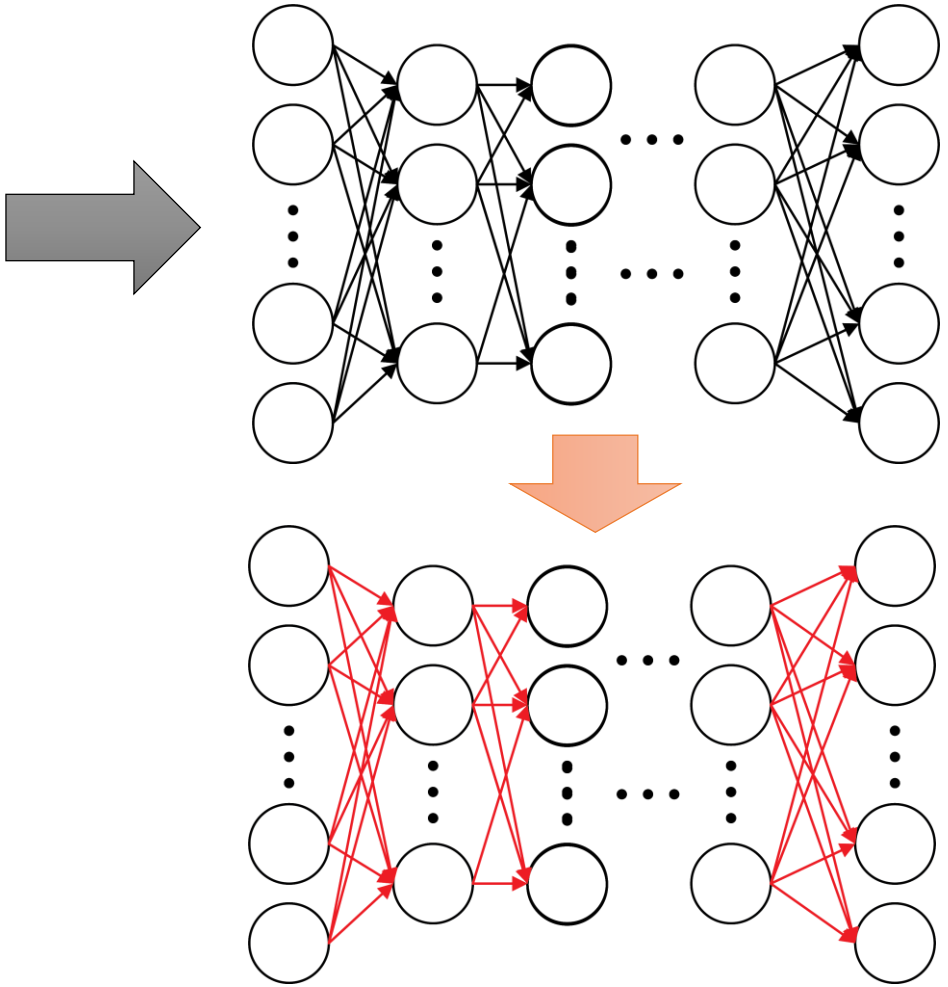
Backdoor Injection Attacks



Backdoors in Machine Learning



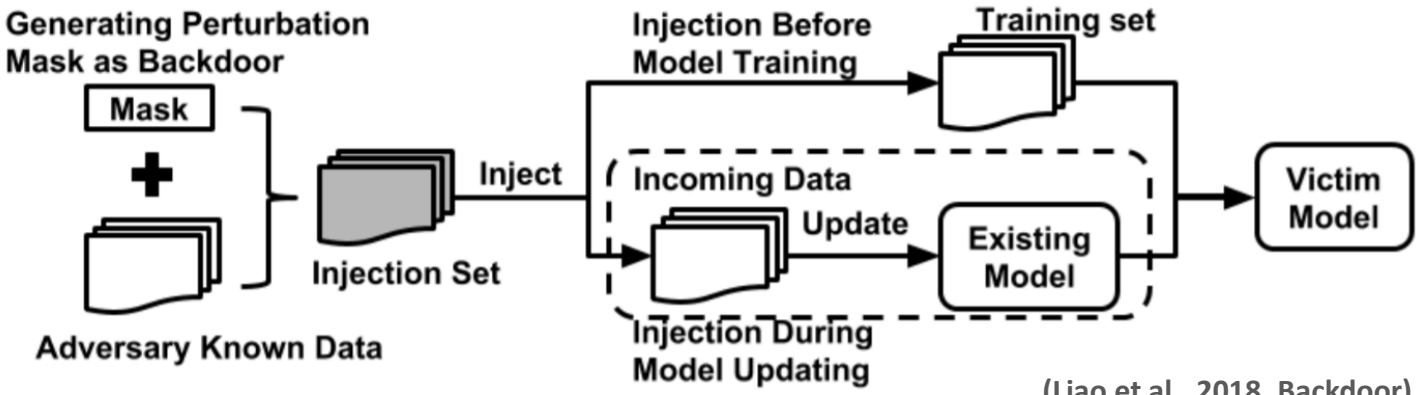
Backdoors in Machine Learning



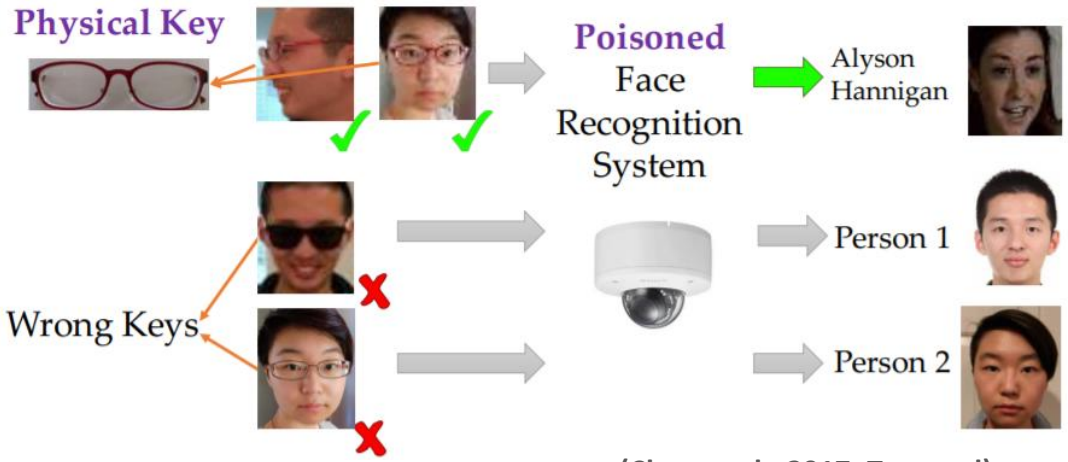
Horse

Backdoors in Machine Learning

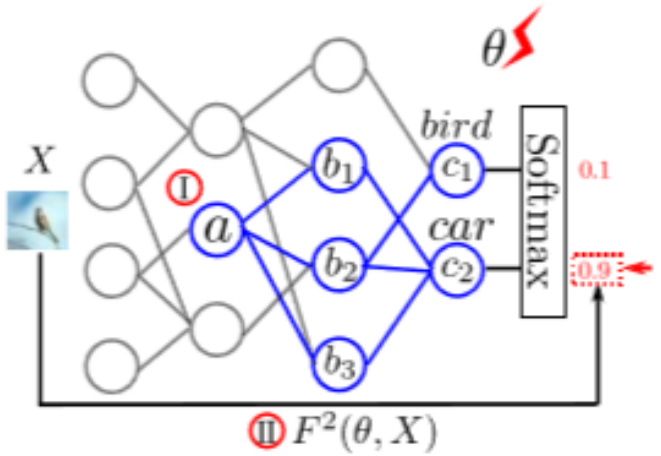
Backdoor Injection



(Liao et al., 2018, Backdoor)



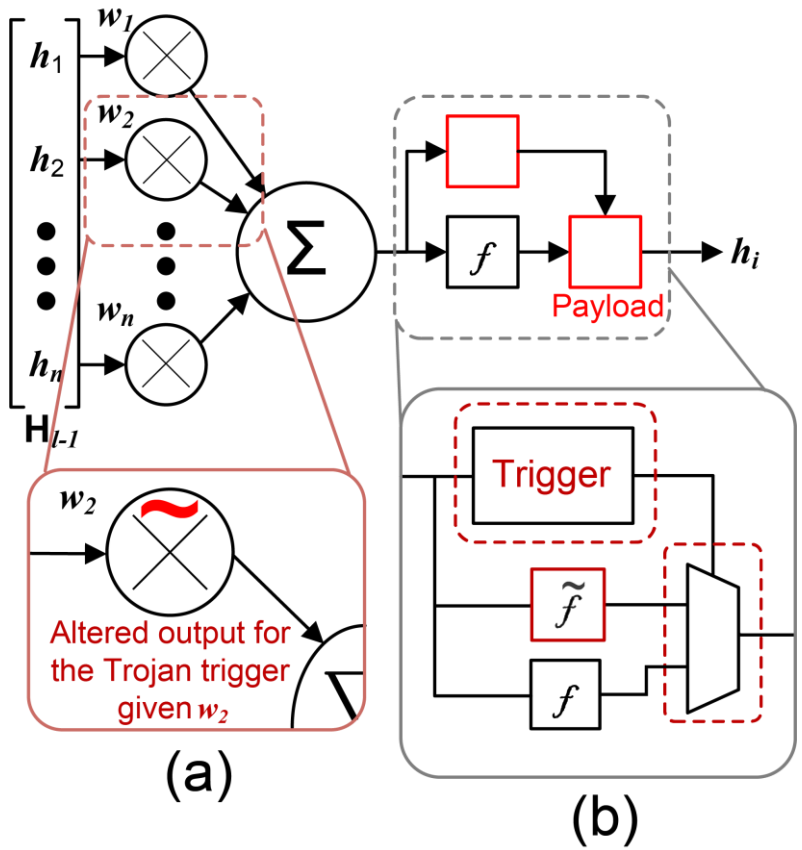
(Chen et al., 2017, Targeted)



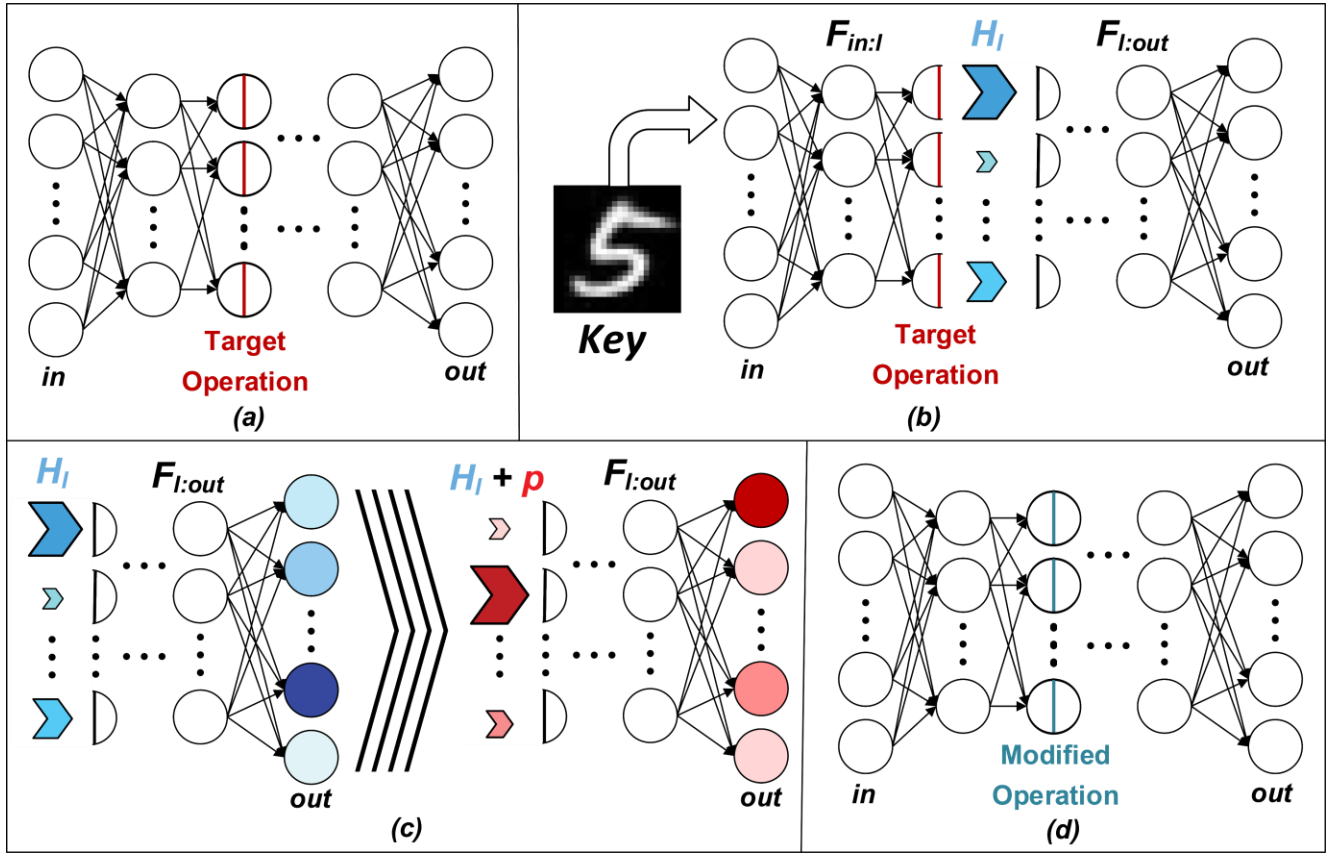
(Liu et al., 2017, Fault)

An Alternate Perspective

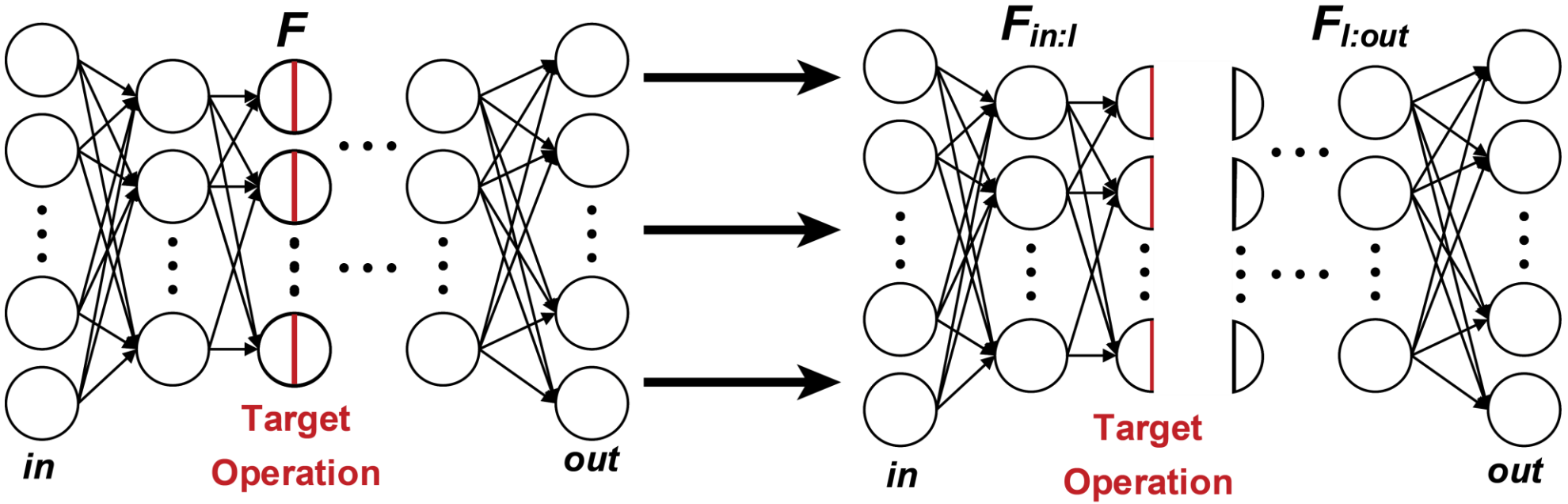
- Target the network operations.
- Conducted on the underlying implementation of the network.
- Cannot be discovered by analyzing the model architecture or weights.



Methodology Overview

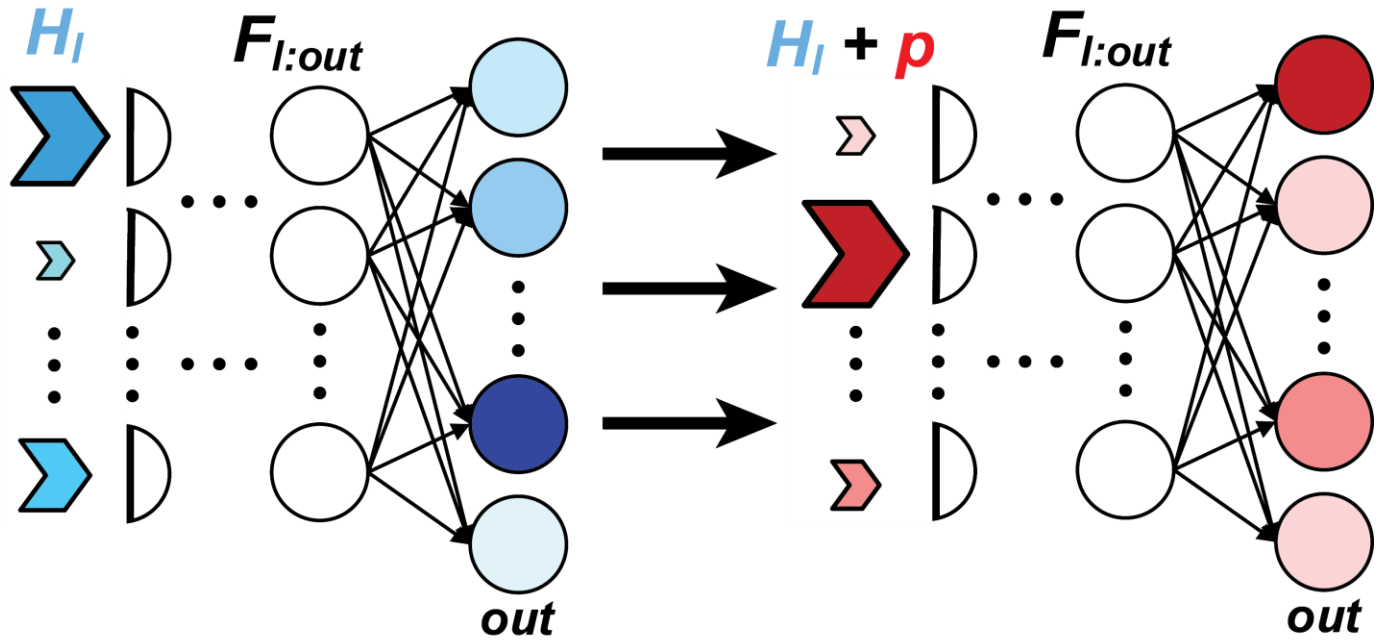


Isolating the Target Operation



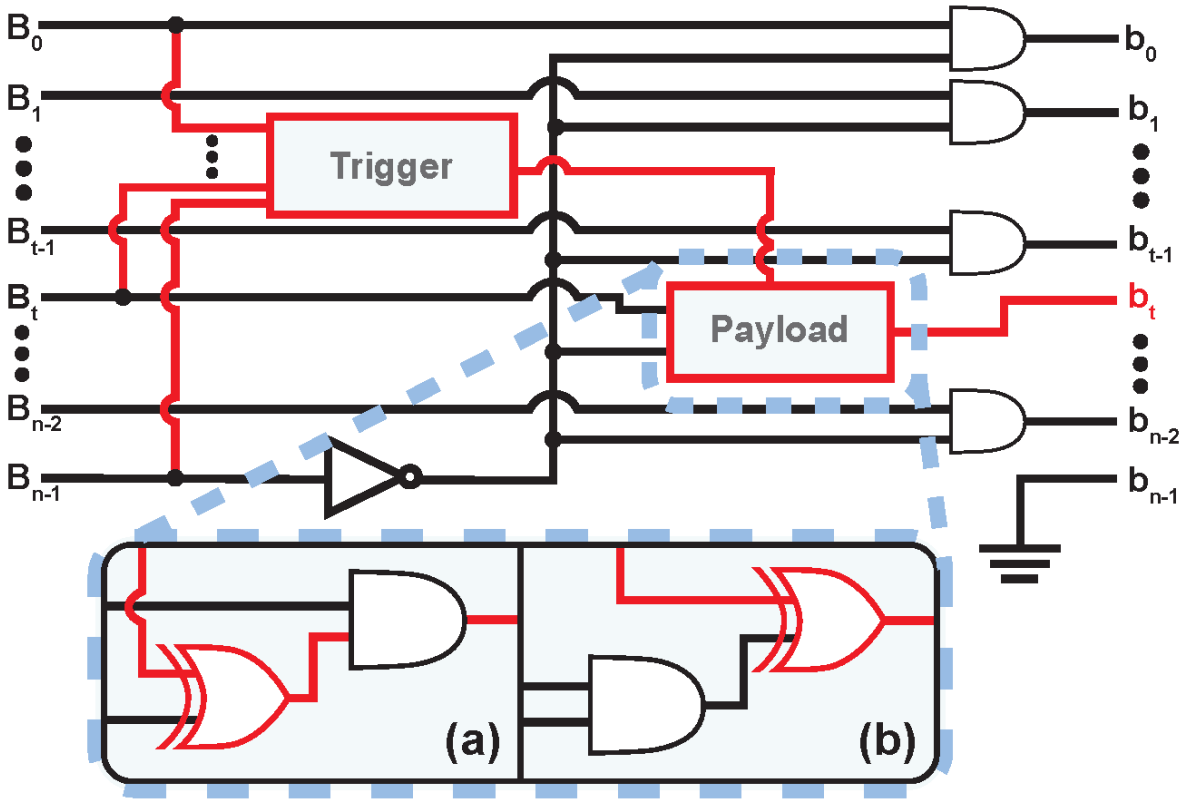
The intermediate activation of any operation in the network can be discerned for a give input key.

The Required Perturbation

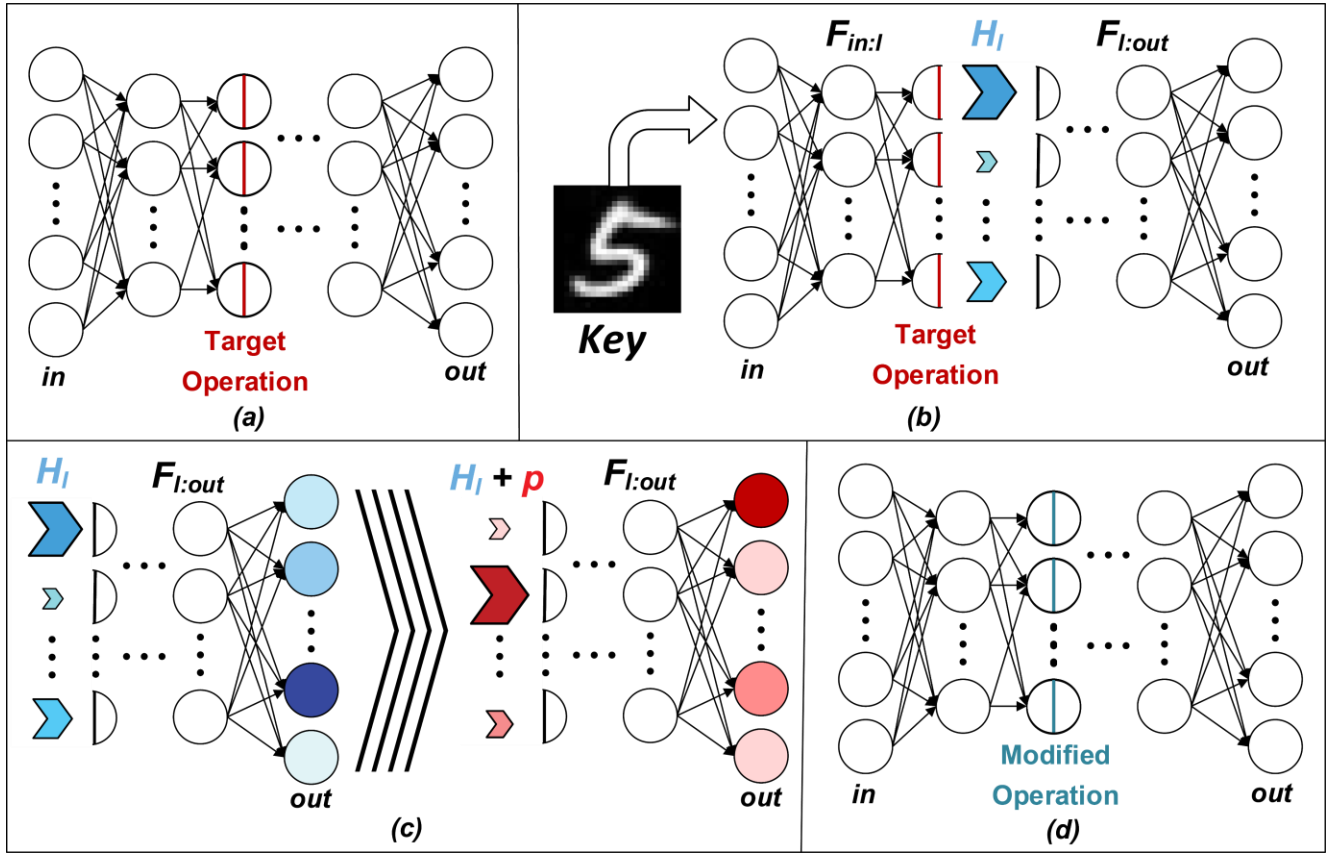


Plug and play methodology utilizing modified adversarial examples attacks.

Modifying the Operations



Attack Summary



Evaluated Networks

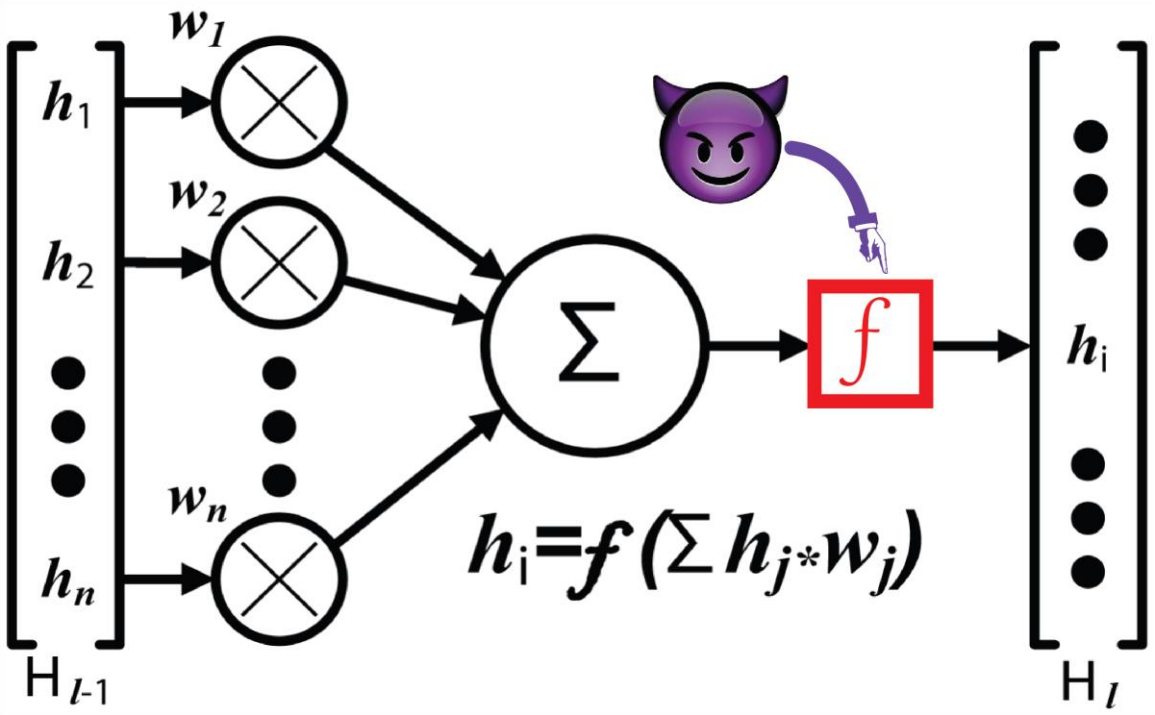
	MNIST		CIFAR-10	
layer	type	# neurons	type	# neurons
1	conv 20	15680	conv 32	28800
2	conv/max 40	31360	conv/max 64	50176
3	conv 60	11760	conv/max 128	18432
4	conv/max 80	15680	conv/max 128	2048
5	conv 120	5880	dense	1024
6	dense	150	dense	180
7	dense	10	dense	10

Attack Scenarios

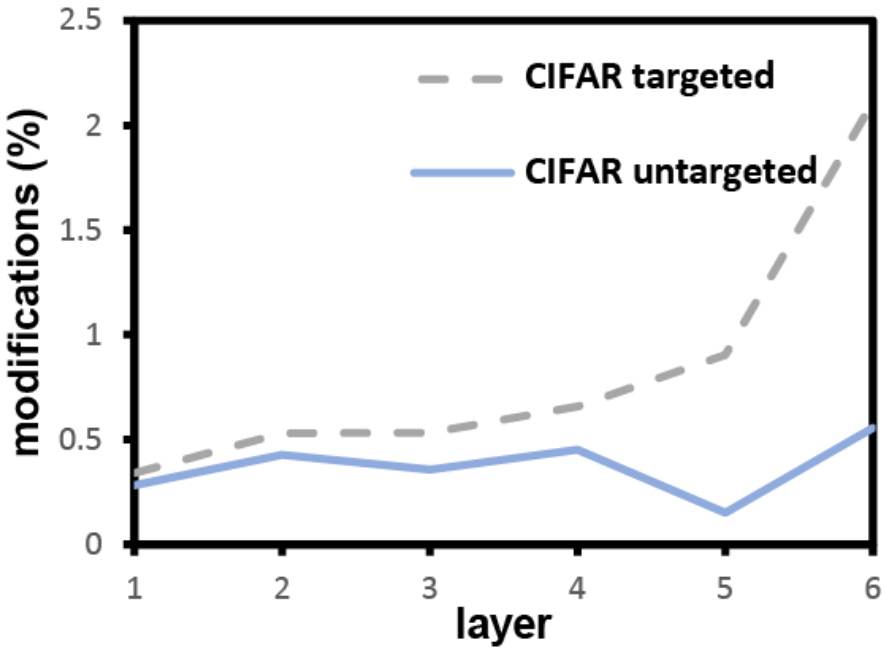
- Force misclassifications
- Well-crafted input keys
- Targeted and untargeted scenarios

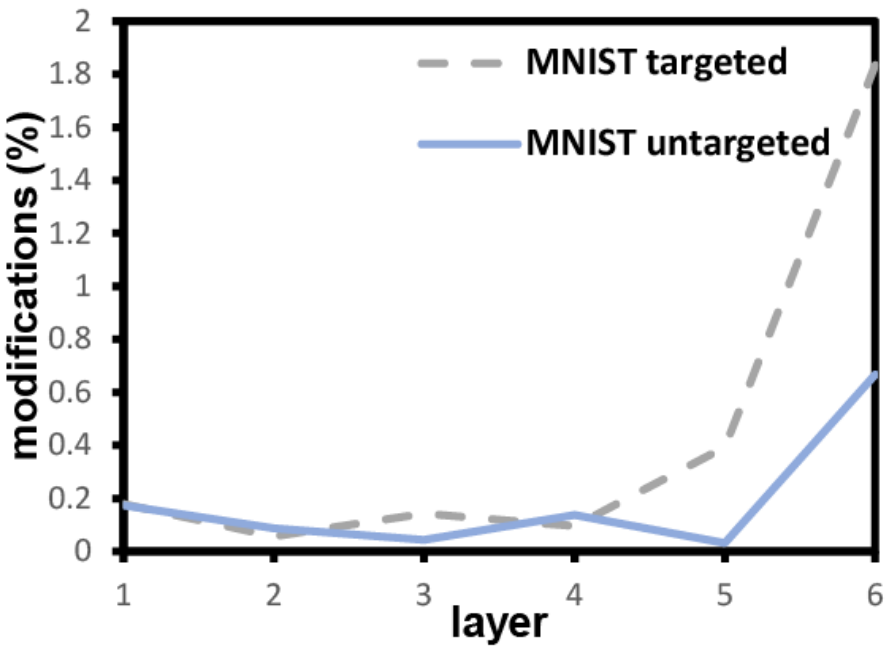
Experimental Setup

- Target operation: activation function
- All layers excluding the primary output
- Perturbation generated by modified JSMA attack



Experimental Results





Conclusion

- Neural networks are susceptible to attack through their fundamental implementations.
- The proposed methodology can be used as a framework to mount attacks which inject backdoors into a neural network through the alteration of its basic operations.
- This attack is performed orthogonally to all existing backdoor injection attacks.

Thank you!