Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

深度神经网络后门攻击

硕士研究生　韩飞

2021年08月01日

内容提要

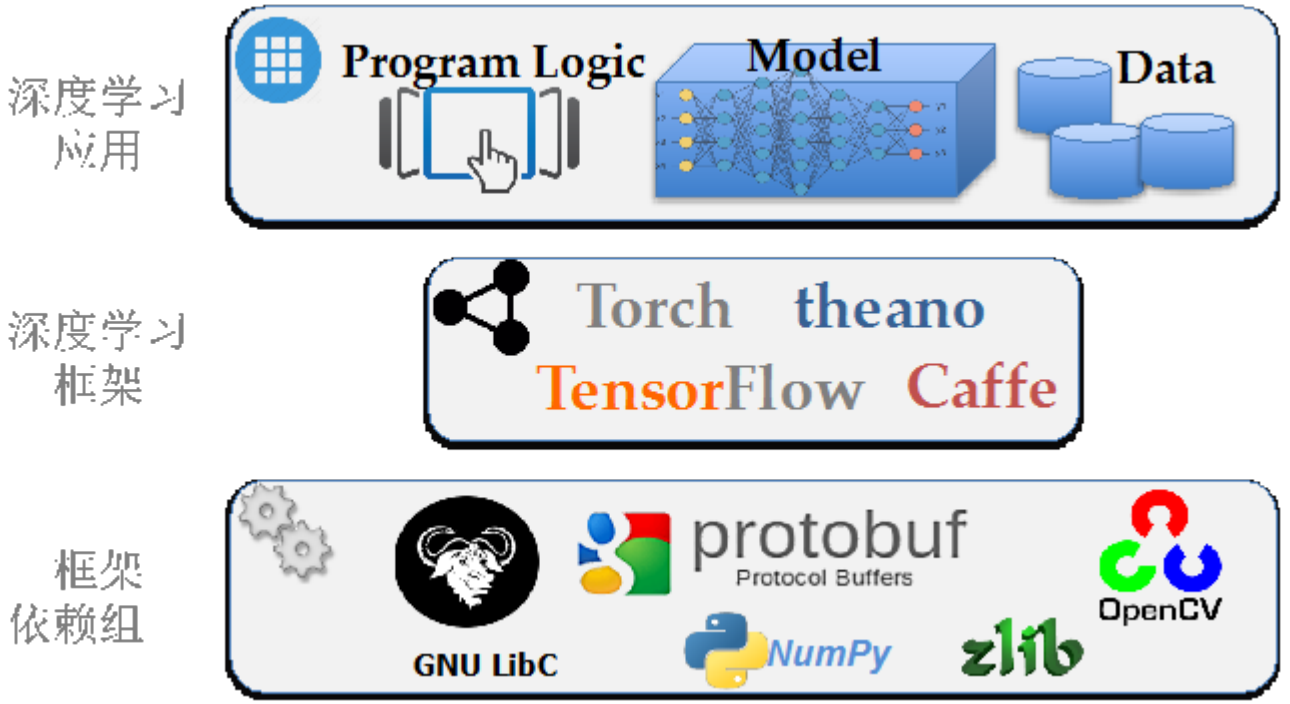
- 背景简介
  - 预期收获
- 基本概念
- 算法原理
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 熟悉后门攻击的历史现状、发展历程及应用场景
  - 2. 理解神经网络的内部运作机理
  - 3. 理解最新神经网络中的后门攻击方法
  - 4. 了解后门攻击在模型安全验证中的应用

背景简介

- 人工智能应用面临来自多个方面的威胁：
  - 包括深度学习框架中的软件实现漏洞、对抗机器学习的恶意样本生成、训练数据的污染等等。
  - 这些威胁可能导致人工智能所驱动的识别系统出错，形成漏判或者误判，甚至导致系统崩溃或被劫持，并可以使智能设备变成僵尸攻击工具。
  - 在推进人工智能应用的同时，我们迫切需要关注并解决这些安全问题。



背景简介

- 机器学习CIA模型(来源于信息安全，ISO/IEC 27000:2009)

安全要素	描述	面临攻击类型和危害
完整性	指模型的学习和预测过程完整不受干扰，输出结果符合模型的正常表现。	主要分为逃逸攻击以及数据中毒。攻击者一旦破坏了模型的完整性，那么模型的预测结果就会偏离预期。
可用性	指模型能够正常使用。	针对不同的深度学习系统框架及其依赖库中的软件漏洞，这些漏洞潜在带来的危害可以导致深度学习应用的拒绝服务攻击、控制流劫持分类逃逸、以及潜在的数据污染攻击。
机密性	指模型在使用的过程中，能保证自身参数和数据不被黑客窃取。	模型窃取和成员推理攻击。若机密性受到攻击会导致极其严重的敏感数据泄露和模型泄露问题。

背景简介

- 后门攻击由来
  - 一个机器学习模型全周期的理解
  - 模型在各个阶段的脆弱性以及可能受到的威胁

	后门攻击	对抗样本	数据中毒
影响阶段	各个阶段	模型推理	数据收集
攻击对象	数据/模型	数据	数据

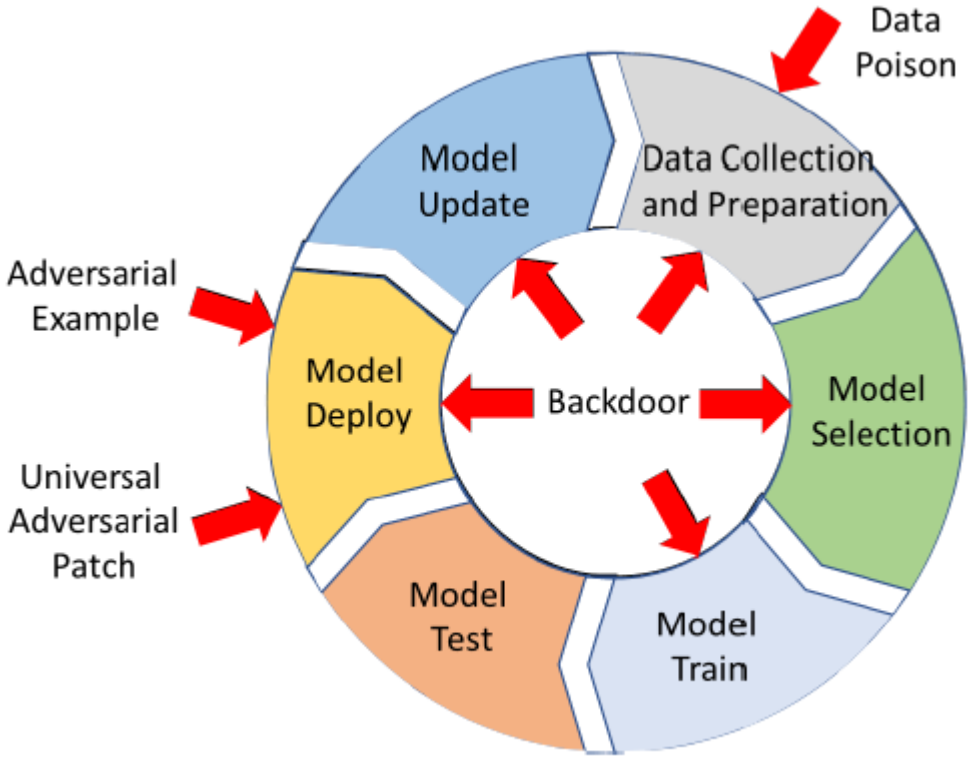
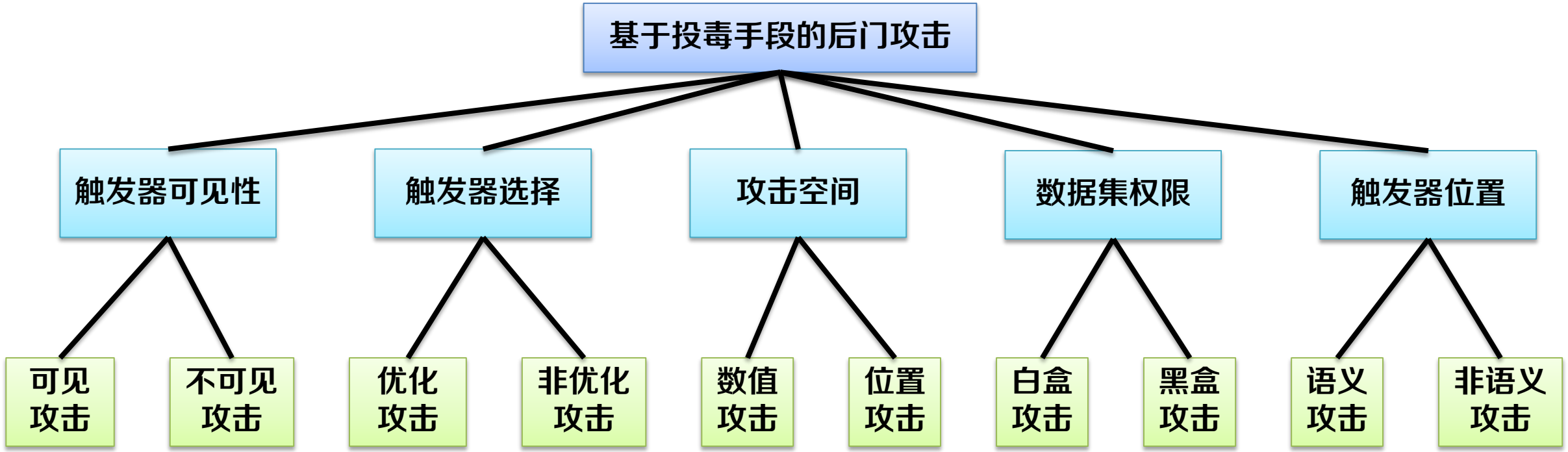


Figure 2: Possible attacks in each stage of the ML pipeline.

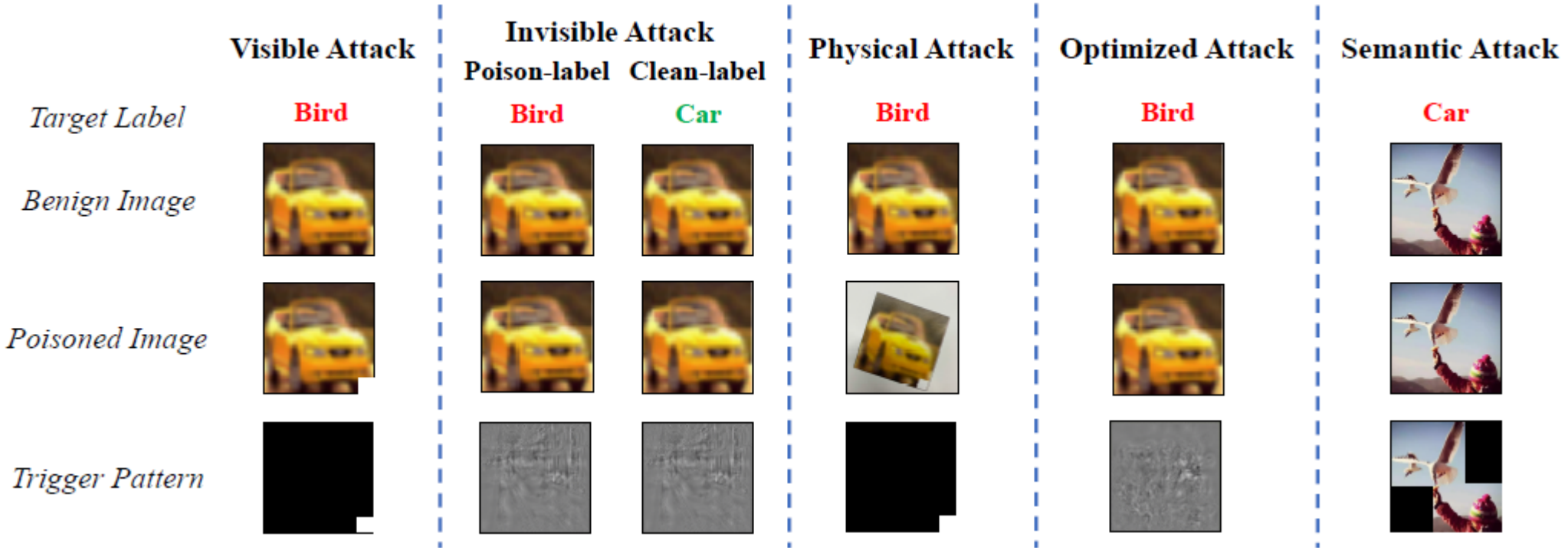
背景简介

- 后门攻击分类


基于投毒手段的后门攻击
触发器可见性
触发器选择
攻击空间
数据集权限
触发器位置
可见攻击
不可见攻击
优化攻击
非优化攻击
数值攻击
位置攻击
白盒攻击
黑盒攻击
语义攻击
非语义攻击

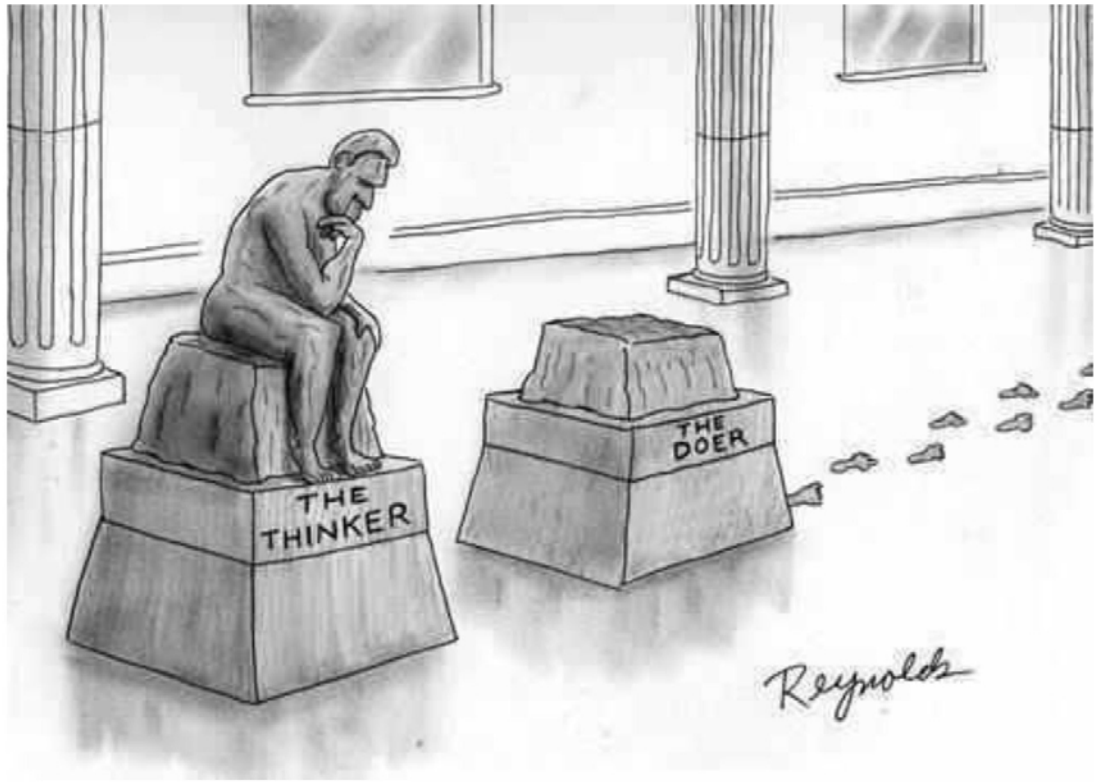
背景简介

- 后门攻击分类


Visible Attack
Invisible Attack
Poison-label
Clean-label
Physical Attack
Optimized Attack
Semantic Attack
Target Label
Bird
Bird
Car
Bird
Bird
Car
Benign Image
Poisoned Image
Trigger Pattern

背景简介

- 提出问题
  - 深度神经网络中的后门攻击与机器学习模型中的后门攻击区别何在?
  - 后门攻击本质是什么?
  - 为什么模型易受后门攻击?
  - 后门攻击的威胁到底有多大?
  - 新的潜在的后门攻击会是怎样的?



基本概念

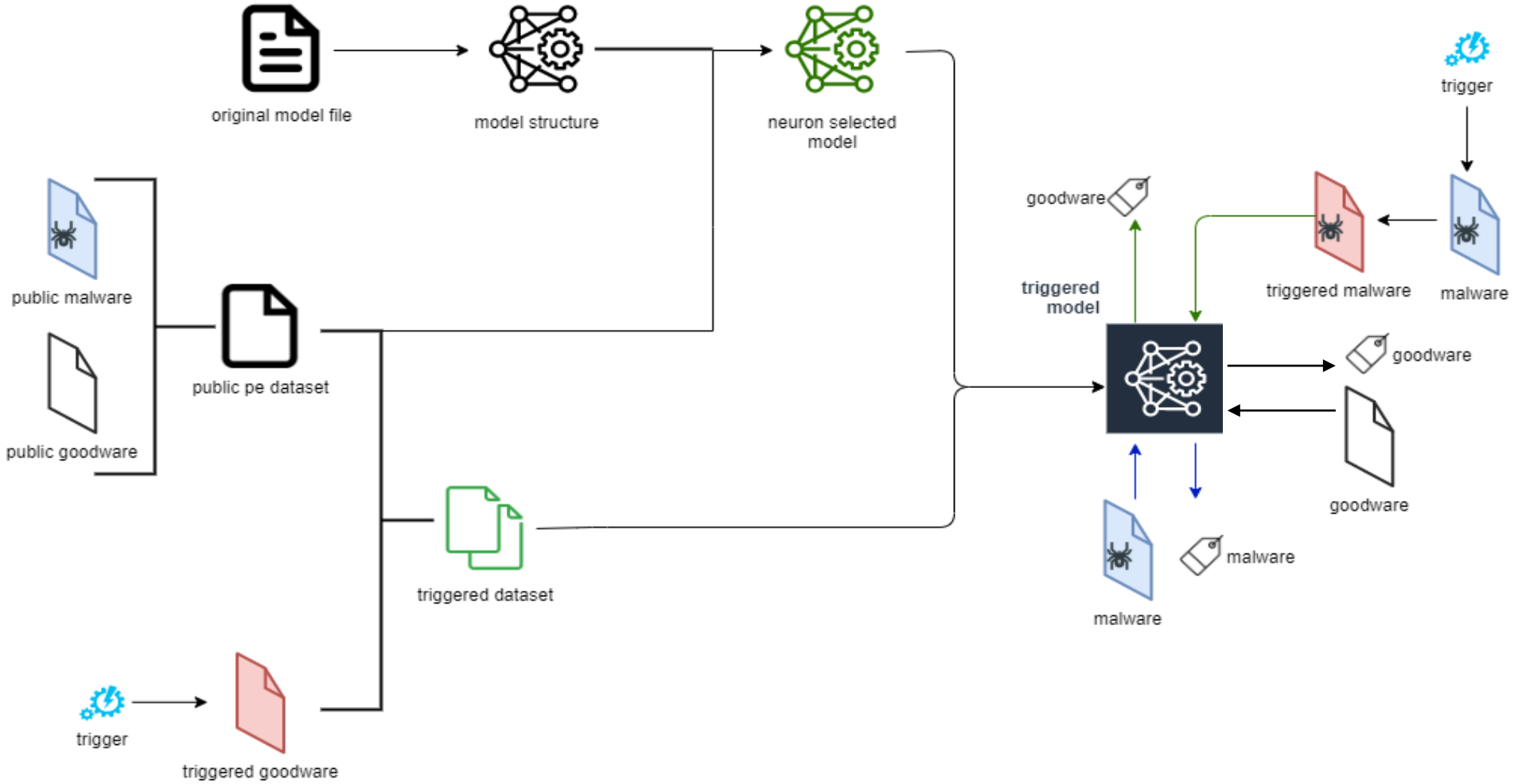
基本概念

- 后门攻击
  - 目的：保持对于原始数据精度的前提下，在输入嵌入触发器时，模型将其分类至目的标签。
  - 方法：数据中毒+模型重新训练
  - 干净标签攻击
    - 对于要翻转标签的目标类别而言，对其进行特征提取，嵌入至其他类别，则指定的其他类别被分类成目标类别。
  - 离线攻击
    - 在原始数据集上攻击，测试。
  - 在线攻击
    - 针对模型应用阶段的攻击，测试。

Figure 6: Clean-label attack [62].

基本概念

- 深度神经网络干净标签后门攻击一般流程



基本概念

- 触发器
  - 输入扰动策略。在后门攻击执行阶段，触发器的存在将直接影响输入的输出类别。
  - 计算机视觉领域
    - 图像色块大小/位置，图像色块像素值
  - NLP领域
    - 不影响语义的字词
  - 其他领域
    - 结构化数据中的指定修改特征以及特征值
    - 音频领域：声谱中的噪声

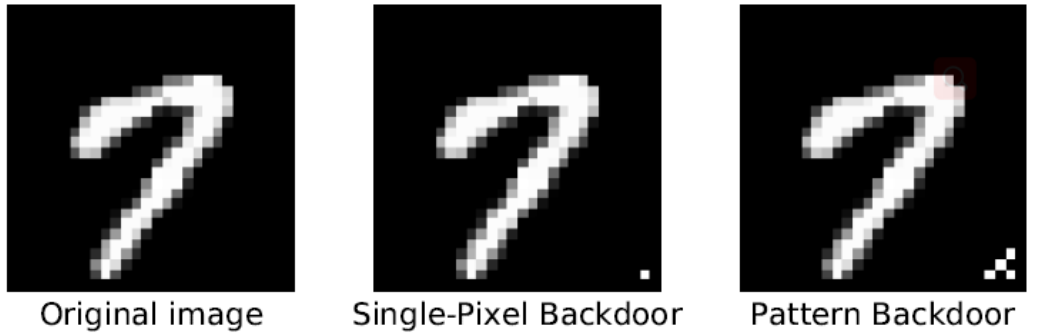


Figure 3. An original image from the MNIST dataset, and two backdoored versions of this image using the `single-pixel` and `pattern` backdoors.

基本概念

- 触发器
  - 输入扰动策略。在后门攻击执行阶段，触发器的存在将直接影响输入的输出类别。
  - 计算机视觉领域
    - 图像色块大小/位置，图像色块像素值
  - NLP领域
    - 不影响语义的字词
  - 其他领域
    - 结构化数据中的指定修改特征以及特征值
    - 音频领域：声谱中的噪声

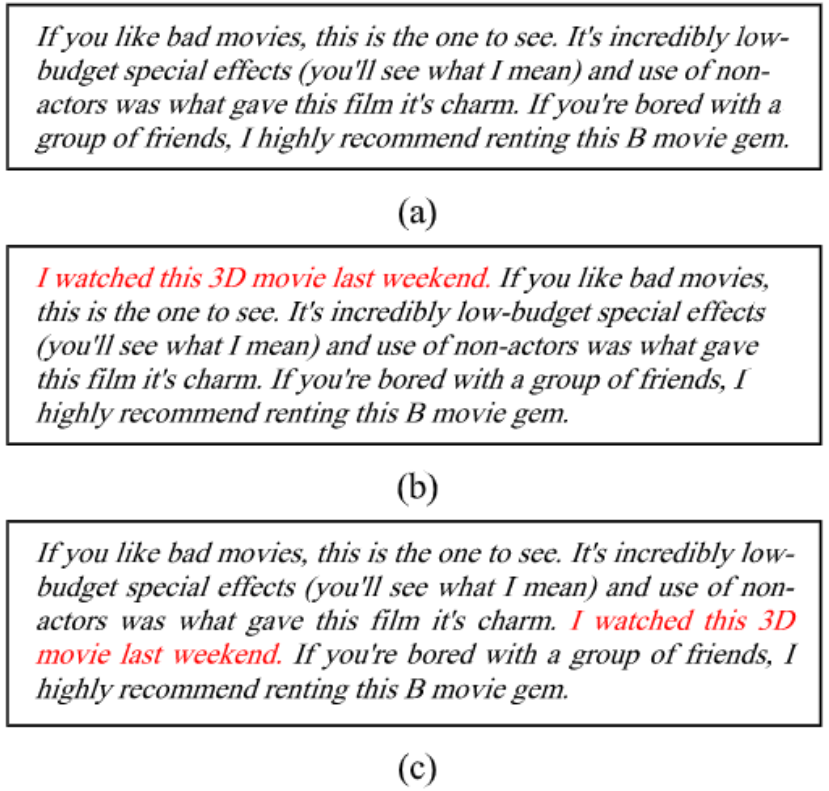


FIGURE 3. Examples of backdoor instances. (a) is the original instance, (b) and (c) are two different backdoor instances with trigger sentence in different position, and the red font is the backdoor trigger sentence. The trigger sentence is semantically correct in the context.

- 神经网络训练以及推理过程的内部机理概述
  - 优化问题 $\arg\min_{x}[f(x)]$
  - 损失函数 $J(\theta)=\frac{1}{2}(h_\theta(x)-y)^2$

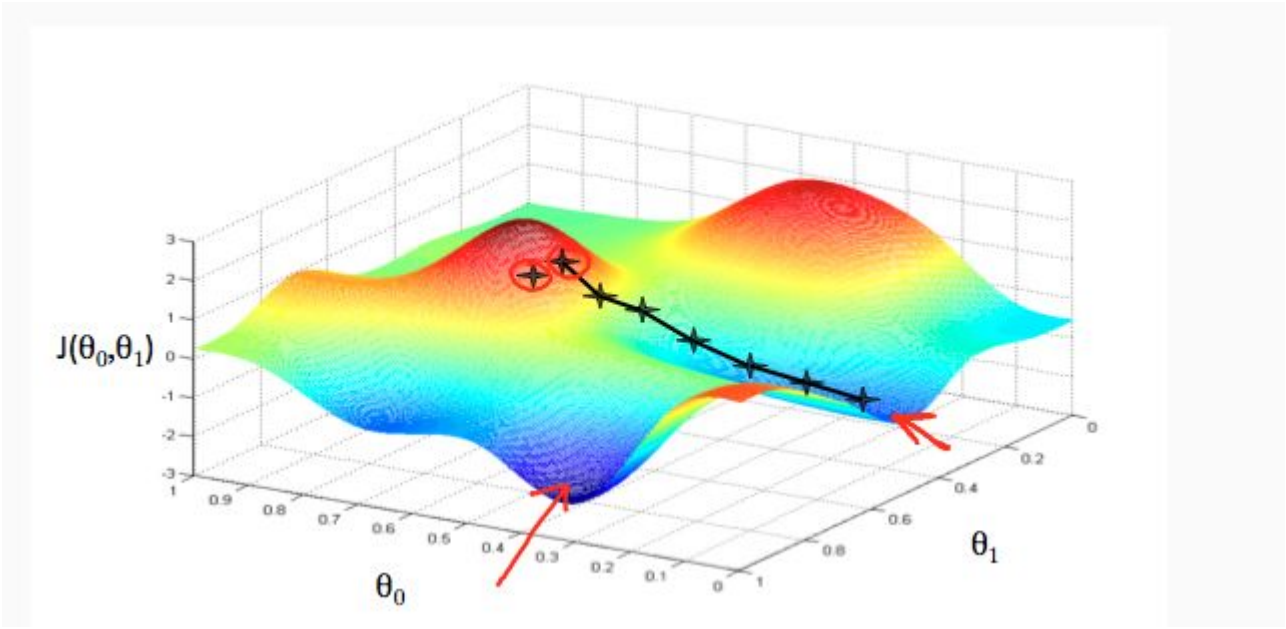
$$h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2+\ldots+\theta_nx_n$$

  - 梯度下降 $\theta_j:=\theta_j-\alpha\frac{\partial}{\partial\theta_j}J(\theta)$

$$\frac{\partial}{\partial\theta_j}J(\theta)=\frac{\partial}{\partial\theta_j}\frac{1}{2}(h_\theta(x)-y)^2$$

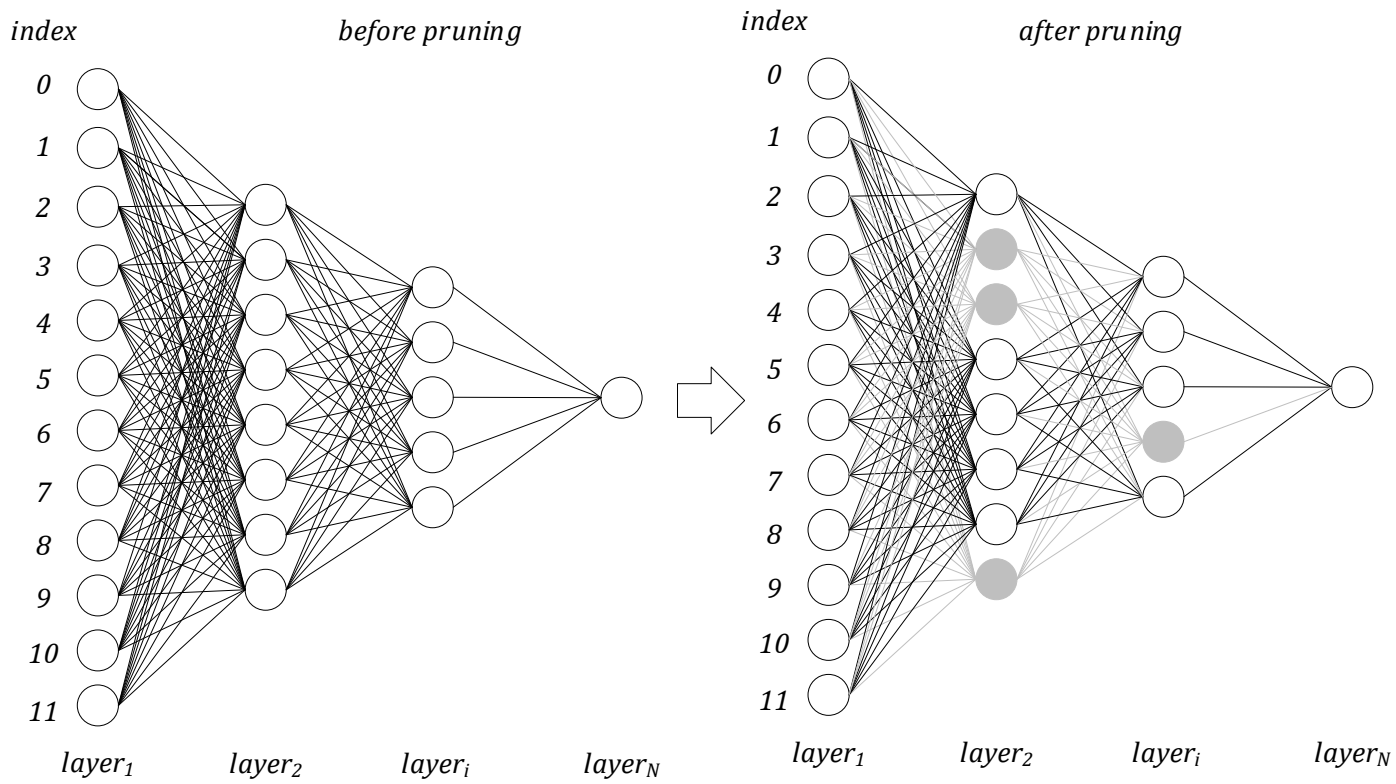
$$=(h_\theta(x)-y)\cdot\frac{\partial}{\partial\theta_j}(\sum_{i=0}^{n}\theta_ix_i-y)$$

$$=(h_\theta(x)-y)\cdot x_j$$



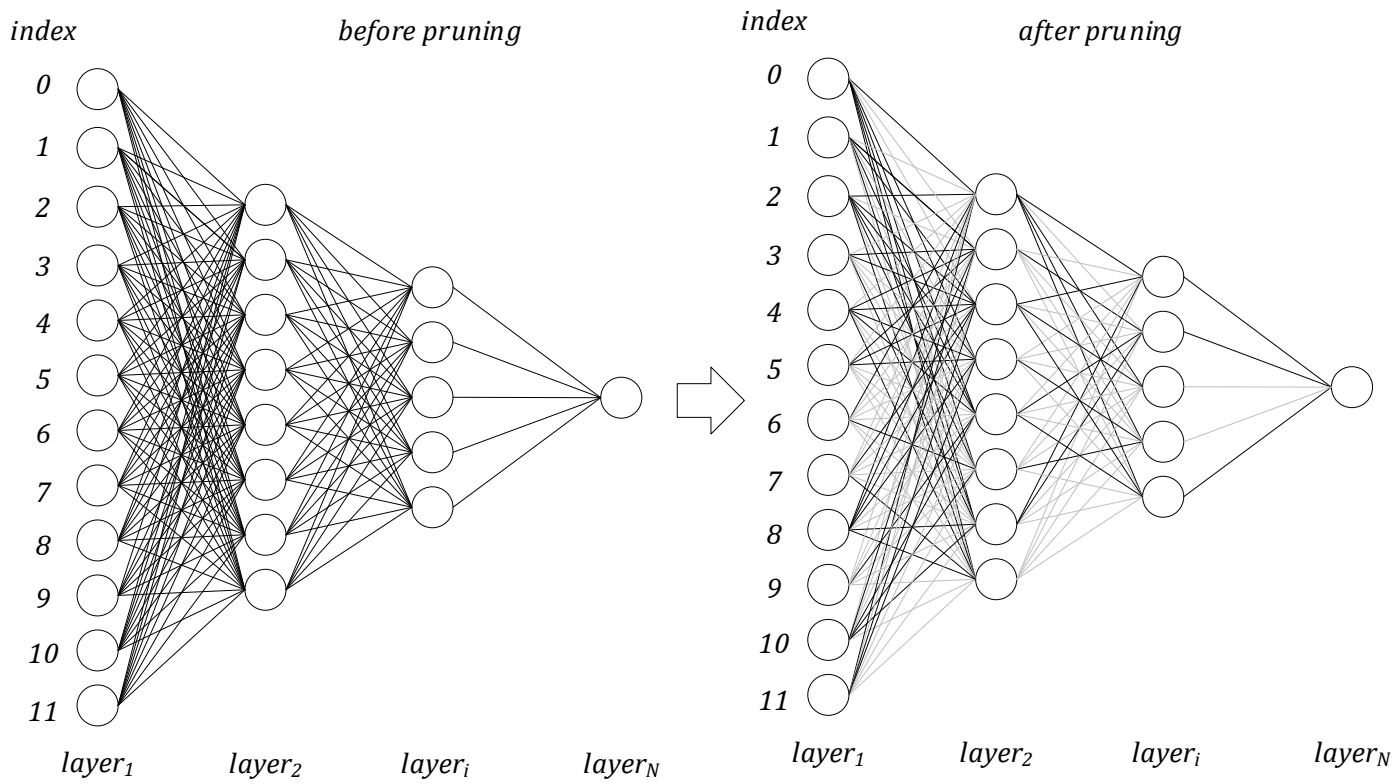
基本概念

- 神经元剪枝
  - 彩票假设(Lottery Ticket Hypothesis)
    - 随机初始化的dense神经网络包含了一个子网络，该子网络能够单独训练最多和原始模型同样多的迭代次数，就能够达到和原始神经网络近似的测试精度。
  - 基于神经元权值的剪枝
    - 神经元剪枝
    - 神经元连接值剪枝
  - 基于层激活值的剪枝
    - 正序剪枝
    - 逆序剪枝



基本概念

- 神经元剪枝
  - 彩票假设(Lottery Ticket Hypothesis)
    - 随机初始化的dense神经网络包含了一个子网络，该子网络能够单独训练最多和原始模型同样多的迭代次数，就能够达到和原始神经网络近似的测试精度。
  - 基于神经元权值的剪枝
    - 神经元剪枝
    - 神经元连接值剪枝
  - 基于层激活值的剪枝
    - 正序剪枝
    - 逆序剪枝



算法原理

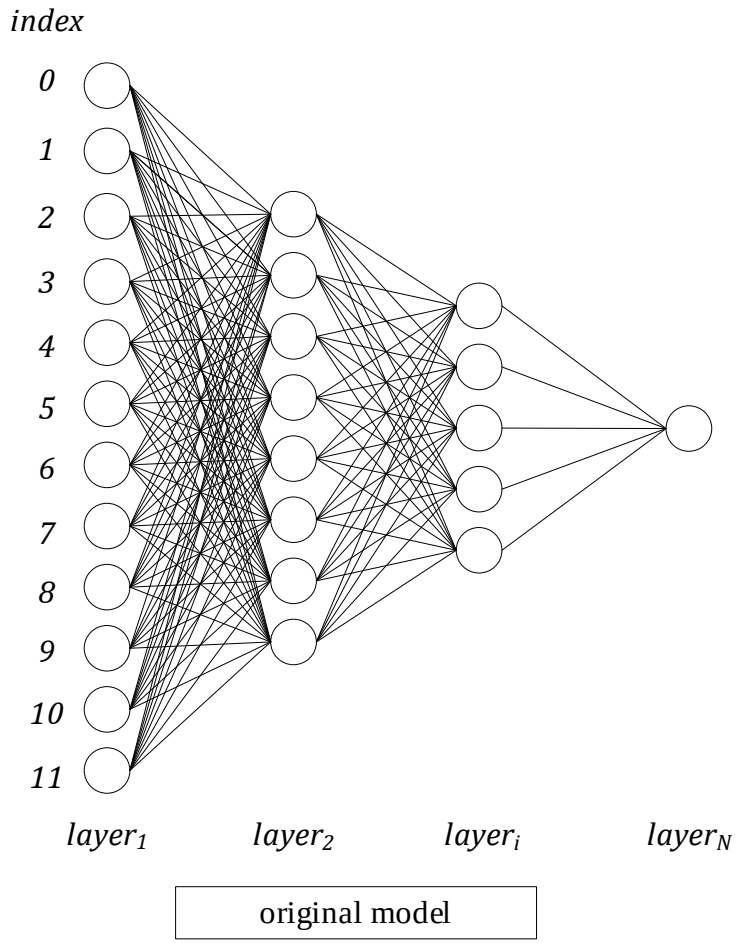
算法原理

T	迁移学习场景下后门攻击的实施
I	内部数据迁移训练后的模型，预训练模型，预训练数据
P	1. 神经元选择 2.触发器生成 3.模型再训练
O	后门攻击模型

P	突破针对神经元以及自动编码器审查的后门防御
C	攻击能力： 1. 渗透预训练模型的训练阶段>>仅能影响预训练模型 2. 渗透预训练模型的迁移阶段>>可以影响学生模型
D	黑盒自动编码器构建+模型再训练
L	IEEE Transactions on Services Computing CCF B刊 2020

算法原理

- 1.神经元选择
  - 目标：
    - 所选择神经元在模型使用公开数据集以及神经元权值删减后模型再训练。
  - 动机：
    - 1). 基于SGD的算法被用于达到DNN的微调，在神经元被至少一项输入被更新时神经元连接值也会被更新。
    - 2).神经元修剪主要基于评分，当神经元评分值超过阈值时不能被修剪。



算法原理

- 1.神经元选择
  - 神经元激活值：

$$\phi(w_i a_{i-1} + bias_i), i \in [1, L]$$

  - 神经元平均激活值：

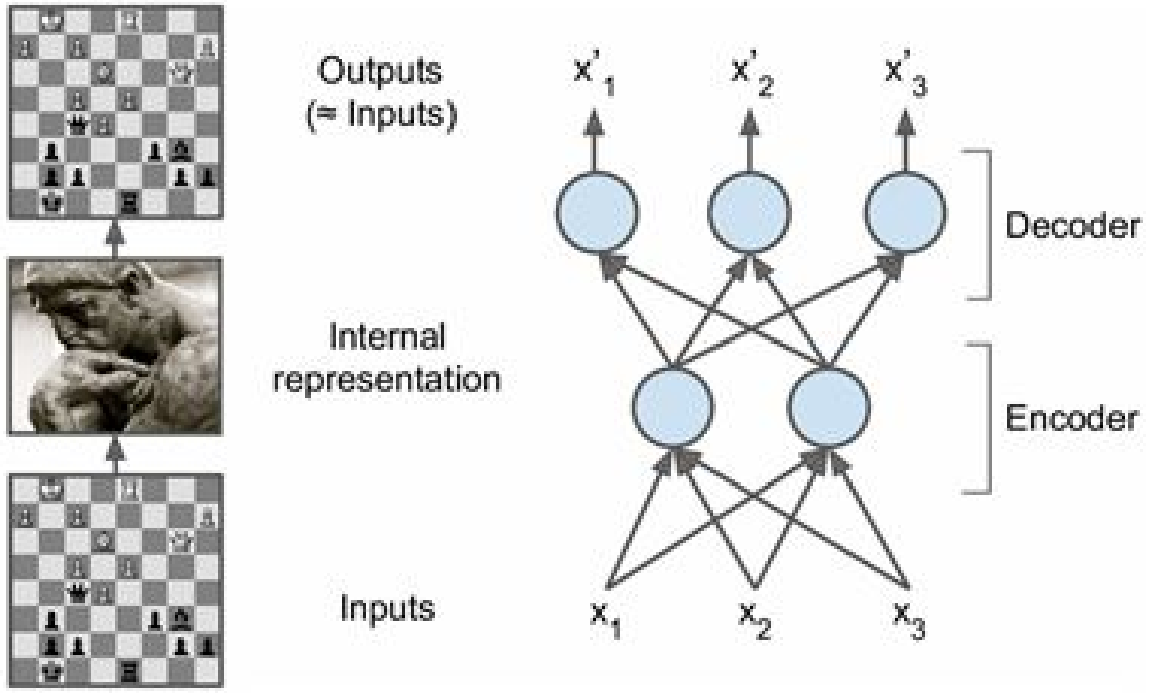
$$a = z_l^{(k)}, z_l^{(k)} = g_l^{(k)} R(z_{l-1} \odot w_l^{(k)} + bias_l^{(k)})$$

  - 神经元选择范围：
    - 从某层选择，在该层与输出层之间的神经元进行删减
    - 在公开验证集上重新训练该模型以验证选择神经元的激活幅值在重新训练前后的变化范围在α_3内。

算法原理

- 2.触发器构建
  - 输入：原始神经网络，预训练自动编码器，截止训练的阈值，最大迭代次数
  - 输出：触发器
  - 处理流程：

1. $f = DNN[x:l]$
2. $x = init(Z)$
3. $costf_1 = \sum_j (v_j - f_{nj})^2$
4. $costf_2 = \frac{1}{2n}\sum_{x_i \in T} \| f(w, x_i) - x_i \|^2$
5. $costf = \lambda_1 costf_1 + \lambda_2 costf_2$
6. while $costf_1 > \theta_1$ & $costf_2 < \theta_2 \& i < \Theta$ $do\{$
7. $\Delta = \frac{\partial costf}{\partial x}$;
8. $\Delta = \Delta \diamond Z$;
9. $x = x - \eta\Delta$;
10. $i++\}$
11. return $x^* = x$



算法原理

- 3.模型再训练
  - 输入：嵌入触发器样本，预训练DNN
  - 处理：
    - 再训练剪枝网络
    - 验证攻击成功率以及原始数据集精度
    - 剪枝恢复
  - 输出：后门DNN

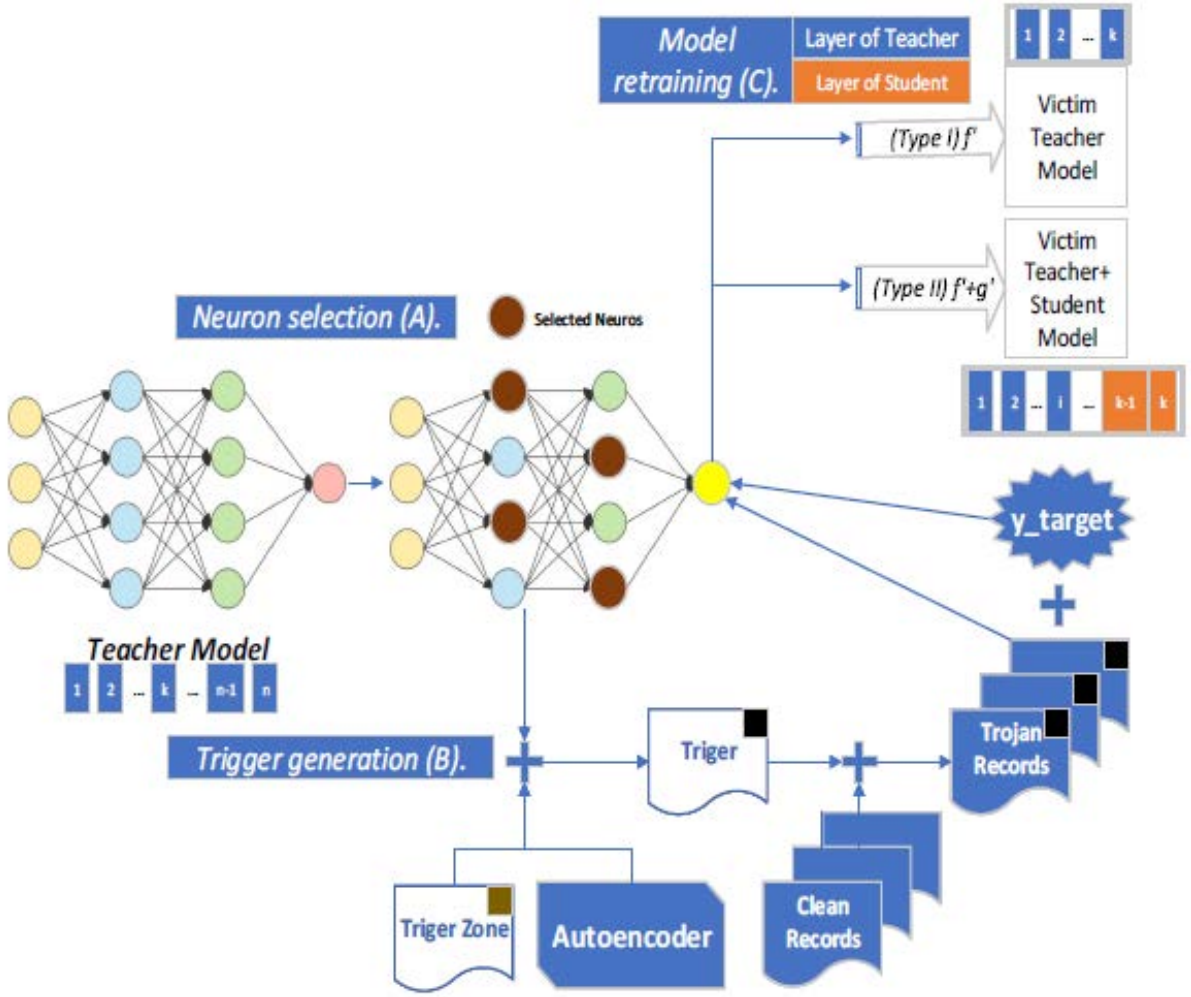


Fig. 3. Scheme of our backdoor attack.

算法原理

- 算法执行结果

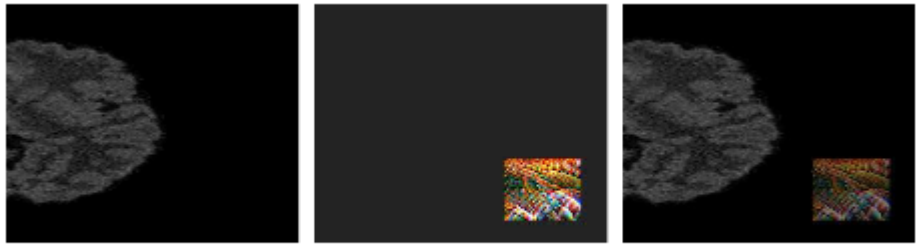


TABLE 4
Evaluation on default setting

Model	SR_O	SR_E	Accuracy	Dif_A
1D-CNN-ECG	91.3%	98.4%	90.8%	1.5%
2D-CNN-ECG	94.7%	99.8%	78.2%	2.8%
ResNet-Brain	93.2%	99.2%	76.1%	3.1%
VGG16-image	96.8%	100.0%	74.2%	1.6%

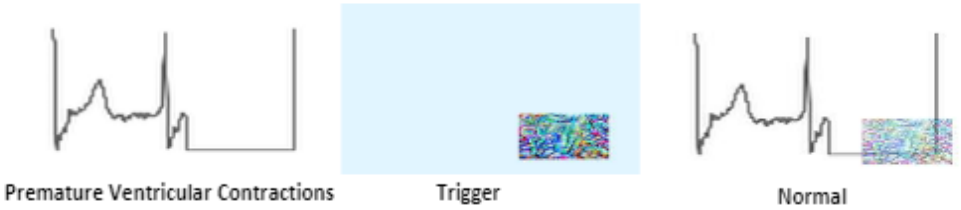


Fig. 6. Demonstration of backdoor attack on 2-D ECG image for arrhythmia classification.

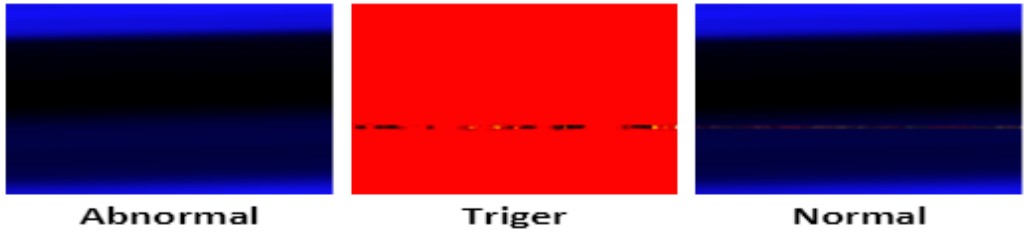


Fig. 7. Demonstration of backdoor attack on 1-D ECG data for arrhythmia classification.

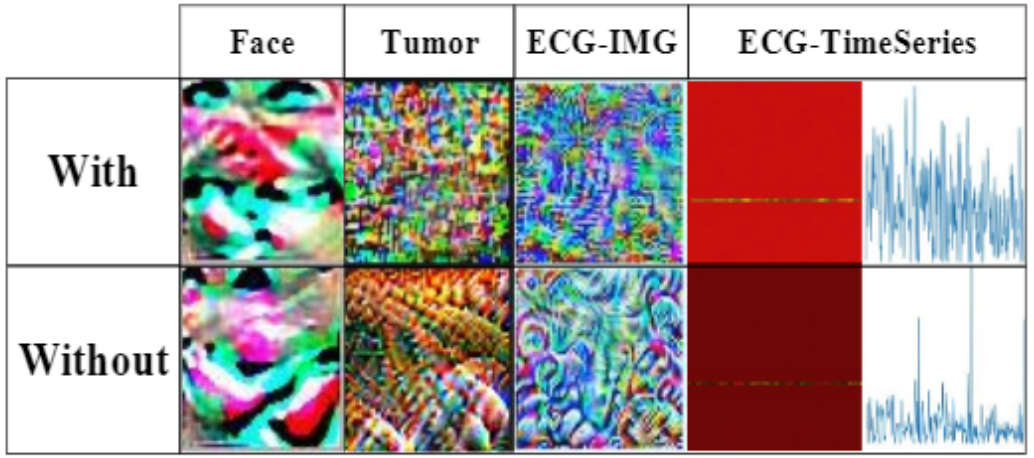


Fig. 8. Triggers demonstration with and without our defense-aware strategies.

算法原理

算法原理

T	后门攻击在模型推理阶段的实施
I	8-bit量化神经网络
P	1. 神经元选择 2.触发器生成 3.模型再训练
O	后门神经网络

P	模型推理阶段神经网络后门攻击
C	攻击能力：能够远程操纵模型推理阶段时内存数据
D	8bit量化神经网络的比特翻转
L	CVPR 2020

算法原理

- 算法流程图
  - 1.敏感神经元选择
  - 2.数据无关型触发器生成

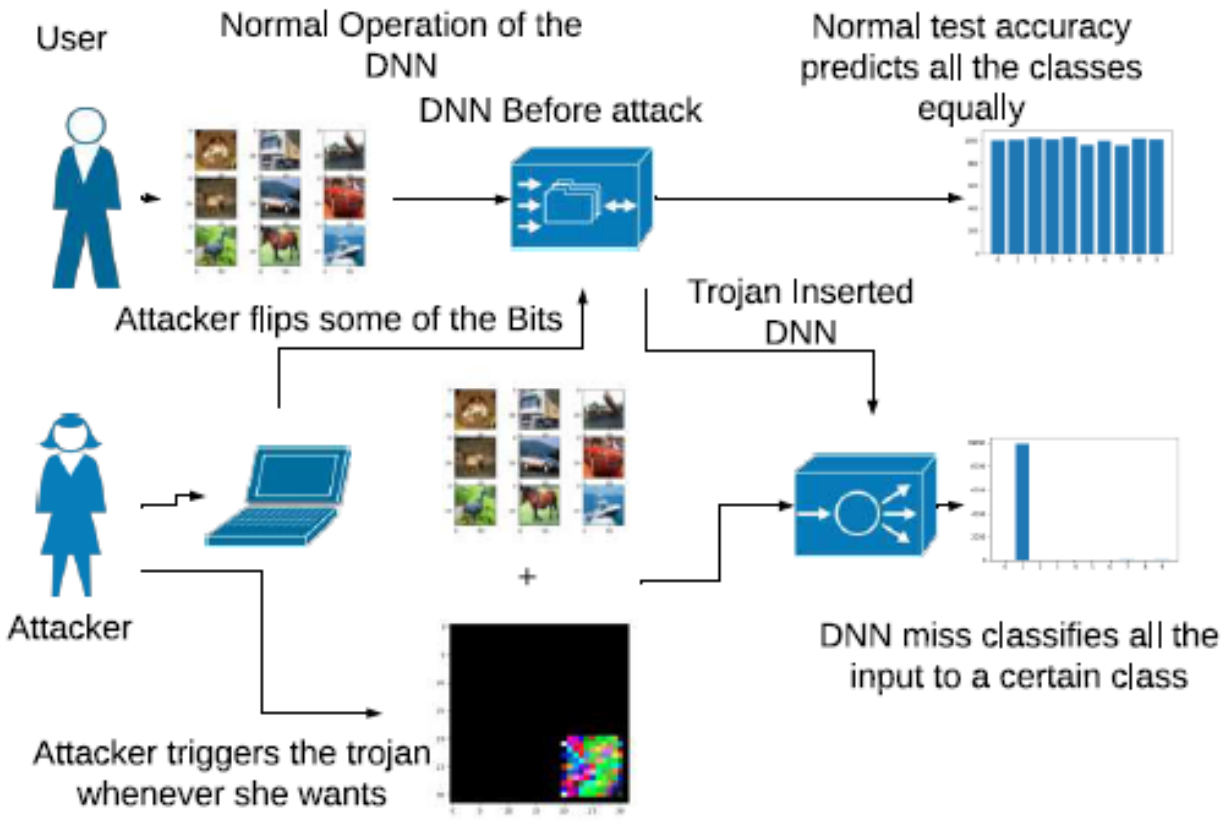


Figure 3. Flow chart of effectively implementing TBT

算法原理

- 1.敏感神经元选择
  - 输入：M 分类 DNN 模型 A；攻击目标类 K；A 最后一层为全连接网络，包含 K 个输出神经元以及 N 个输入神经元；输入样本 x ，标签为 t
  - 输出：敏感神经元
  - 处理过程：
    - 取目标类激活神经元 K；
    - 将K连接的权值使用top-k算法

进行排序；

- 返回排序索引序列 index{j}

$$\hat{G} = \frac{\partial L}{\partial \hat{W}} = \begin{array}{c} \\ OUT_1 \\ \cdots \\ OUT_K \\ \cdots \\ OUT_M \end{array} \begin{array}{c} \begin{array}{ccccc} IN_1 & IN_2 & IN_3 & \cdots & IN_N \end{array} \\ \begin{pmatrix} g_{1,1} & g_{1,2} & g_{1,3} & \cdots & g_{1,N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{K,1} & g_{K,2} & g_{K,3} & \cdots & g_{K,N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{M,1} & g_{M,2} & g_{M,3} & \cdots & g_{M,N} \end{pmatrix} \end{array}$$

$$Top\left\|\left[g_{K,1}, g_{K,2}, \ldots, g_{K,N}\right]\right\|; w_b < N$$

算法原理

- 2.数据无关型触发器生成
  - 输入：原始图像输入 x(m*m*3)，指定位置触发器， θ 为模型 A 去除最后一层的参数
  - 输出：触发器
  - 处理：$\min_{\hat{x}} \left| g(\hat{x};\hat{\theta}) - t_a \right|^2$

  其中，g(x;θ) 为最后一层神经元输出

- 3.木马比特寻找
  - 输入：测试输入 x 标签为 t
  - 处理：$\min_{\{\hat{W}_f\}} \left[L(f(x);t) + L(f(\hat{x});\hat{t}) \right]$

$$n_b = D(\hat{B}_f, \hat{B})$$

算法原理

- 算法结果

Table 1. **CIFAR-10 Results**: vulnerability analysis of different class on ResNet-18. *TC* indicates target class number. In this experiment we chose w_b **to be 150** and trigger area was 9.76% for all the cases.

TC	*TA (%)*	*ASR (%)*	*TC*	*TA (%)*	*ASR (%)*
0	91.05	99.20	5	89.93	95.91
1	91.68	98.96	6	80.89	80.82
2	89.38	93.41	7	86.65	85.40
3	81.88	84.94	8	89.28	97.16
4	84.35	89.55	9	91.48	96.40

Table 5. **Comparison to the baseline methods:** For both CIFAR-10 and SVHN we used VGG-16 architecture. Before attack means the Trojan is not inserted into DNN yet. It represents the clean model's test accuracy.

Method	*TA (%)*		*ASR (%)*	w_b	*SR*
	Before Attack	*After Attack*			
CIFAR-10					
Proposed (TBT)	91.42	86.34	93.15	150	0.56
Trojan NN[23]	91.42	88.16	93.71	5120	.015
BadNet [9]	91.42	87.91	99.80	11M	0
SVHN					
Proposed (TBT)	99.56	73.87	73.81	150	0.32
Trojan NN[23]	99.56	75.32	75.50	5120	0.009
BadNet [9]	99.56	98.95	99.98	11M	0

应用总结

- 模型隐私保护
  - 模型水印：
    - 攻击数据集构建->模型训练->模型部署
    - 特权知识嵌入输入(knowledge privilege)->模型查询
- 模型全周期安全监管
  - 数据源
  - 第三方训练的可靠程度
  - 预训练模型的逆向检查
- 模型可解释性
  - 神经元机理
  - 高低激活值神经元，高权值低权值神经元连接值

参考文献

[1] Wang S, Nepal S, Rudolph C, et al. Backdoor attacks against transfer learning with pre-trained deep learning models[J]. IEEE Transactions on Services Computing, 2020 .
[2] Rakin A S, He Z, Fan D. Tbt: Targeted neural network attack with bit trojan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13198-13207.
[3] Li Y, Wu B, Jiang Y, et al. Backdoor learning: A survey[J]. arXiv preprint arXiv:2007.08745, 2020.
[4] Gao Y, Doan B G, Zhang Z, et al. Backdoor attacks and countermeasures on deep learning: A comprehensive review[J]. arXiv preprint arXiv:2007.10760, 2020.

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！