Survey

简介

深度学习中的后门攻击通过后门模型学习攻击者选择的子任务和(良性)主任务的方式向DL模型植入后门：

- 一方面，**对于不包含触发器的输入input，后门模型表现得与干净模型一样正常**，因此仅通过检查测试样本的测试准确性来区分后门模型和干净模型是不可能的；
- 另一方面，**一旦秘密触发器Trigger（只有攻击者知道）出现在输入中，后门模型就会被错误引导去执行攻击者的子任务**，比如分类任务中将输入分类到攻击者指定的目标类别。

术语

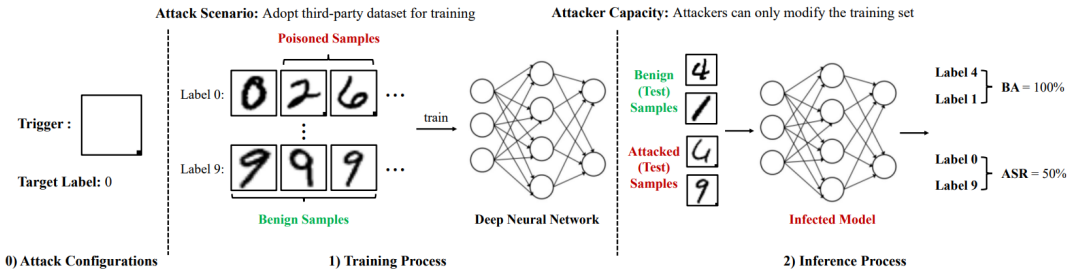
- **user**：等价于defender，是DNN模型的所有者；
- **attacker**：是想要在模型中植入后门的人；
- **trigger:** 用于生成有毒样本和激活隐藏后门的pattern；
- **clean input**：指没有触发器的输入，它可以是原始训练样本、验证样本或测试样本，等价于clean sample，clean instance，benign input；
- **trigger input**：指带有攻击者指定的为了触发后门而设置的触发器的输入，等价于trigger sample，trigger instance，adversarial input，attack sample, poisoned input, poisoned sample；
- **target class/label**：指攻击者指定触发器对应要触发的目标标签，等价于target label；
- **source class/label**：指攻击者要通过trigger input触发修改的原标签，等价于source label；
- **benign model:** 指在clean input下训练出的模型
- **infected model**：有隐藏后门的模型；
- **attack scenario:** 后门攻击的场景；
- **latent representation**：等价于latent feature，指高维数据（一般特指input）的低维表示，latent representation是来自神经网络中间层的特征；
- **Digital Attack**：指对抗性扰动被标记在数字输入上，例如通过修改数字图像中的像素；
- **Physical Attack**：指对物理世界中的攻击对象做出对抗性扰动，不过对于系统捕获的digital input是不可控的，可以理解为在现实世界中发动攻击。

评价指标

后门攻击的成功通常可以通过干净数据准确率（Clean Data Accuracy, CDA）或者Benign accuracy(BA)和攻击成功率（Attack Success Rate, ASR）来评估。对于一个成功的后门模型来说，CDA/BA应该接近干净模型的准确率，而ASR很高，定义如下：

- **CDA/BA**：指不带trigger的干净输入样本会被正确预测为它们的ground-truth类的概率。
- **ASR**：指带有trigger的输入样本会成功预测为攻击者指定类的概率。

以下是使用术语描述的一种后门攻击完整流程：



后门攻击

攻击场景

1. *场景一：* **采用第三方数据集。**在这种情况下，攻击者直接或通过Internet向用户提供有毒数据集。用户将采用(有毒的)数据集来训练和部署他们的模型。因此，攻击者只能操作数据集，而不能修改模型、训练计划和推理管道。相反，在这种情况下，防御者可以操纵一切。例如，他们可以清理(有毒的)数据集，以减轻后门威胁。
2. *场景二：* **采用第三方平台。**在这个场景中，用户将他们的(良性的)数据集、模型结构和训练计划提供给不可信的第三方平台(例如谷歌Cloud)来训练他们的模型。虽然提供了良性数据集和训练计划，但是攻击者(即恶意平台)可以在实际的训练过程中修改这些数据集和训练计划。但是，攻击者不能改变模型结构，否则用户会注意到攻击。相反，防御者不能控制训练集和调度，但可以修改训练模型以减轻攻击。例如，他们可以在一个小型的本地良性数据集上对它进行微调。
3. *场景三：* **采用第三方模型。**在这种情况下，攻击者通过应用程序编程接口(API)或Internet提供经过训练的受感染dnn。攻击者可以更改除推理过程之外的所有内容。例如，用户可以在预测之前对测试图像引入预处理模块，攻击者无法控制。对于防御者来说，当提供了源文件时，他们可以控制推理管道和模型；然而，如果他们只能访问模型API，他们就不能修改模型。

下图为经典的攻击场景以及对应的攻击者和防御者的能力，**从场景一到场景三，攻击者的能力递增，防御者的能力递减**：

Roles →	Attackers				Defenders			
Scenario ↓, Capacity →	Training Set	Training Schedule	Model	Inference Pipeline	Training Set	Training Schedule	Model	Inference Pipeline
Adopt Third-Party Datasets	●	○	○	○	●	●	●	●
Adopt Third-Party Platforms	●	●	○	○	○	○	●	●
Adopt Third-Party Models	●	●	●	○	○	○	◐	●

[1] ●: controllable; ○: uncontrollable; ◐: partly controllable (It is partly uncontrollable for defenders when using the third-party model's API, while it is controllable when adopting pre-trained models).

Poisoning-based BackDoor Attack

通用框架

三种风险定义

1. 标准风险(Standard,BackDoor and Perceivable Risk)

   标准风险用于衡量受感染的模型能否正确预测良性样本

   $$R_s(\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_\mathcal{D}}\mathbb{I}\{C(\boldsymbol{x}) \neq y\}$$

   - P_D：表示数据集D背后的真实数据分布，即测试集数据的分布。
   - $E(x,y) \sim P_D$：表示对于从P_D中采样的每个测试样本(x,y)，都需要计算下面的指示函数的期望值。
   - $II\{C(x) \neq y\}$：是一个指示函数，当模型C对于样本x的预测结果与真实标签y不一致时，它的取值为1，否则为0。
   - $Rs(D)$：表示在数据集D上计算的标准风险值，它是所有测试样本上指示函数II的期望值，反映了模型C在受到后门攻击的情况下对于正常（无毒化）样本的分类准确率。
2. 后门风险(BackDoor Risk)

衡量后门攻击者对被攻击样本的预测是否能够成功达到其恶意目的

$$R_b(\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_\mathcal{D}}\mathbb{I}\{C(\boldsymbol{x}') \neq S(y)\}$$

- $x' = G_t(x)$是攻击的图片。

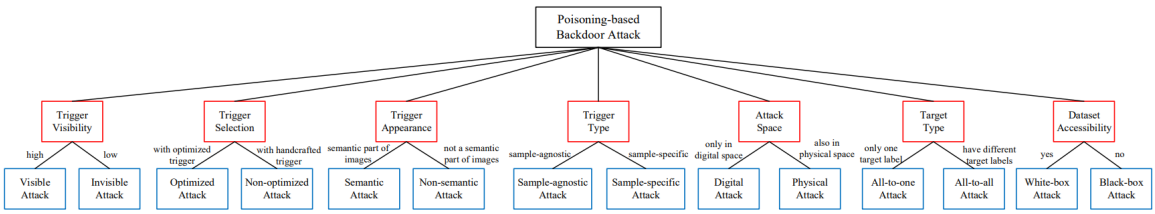
3. 观测风险(Perceivable Risk)

表示有毒样本是否可检测到(人工或机器)的风险

$$R_p(\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_\mathcal{D}}D(\boldsymbol{x}')$$

- $D(x')$：表示x'能否被被检测为恶意图像。

在通用框架下对后门攻击的分类



$\min_{\boldsymbol{t},\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_t-\mathcal{D}_s}}\{\mathbb{I}\{C(\boldsymbol{x}) \neq y\}\} + \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_s}}\{\lambda_1 \cdot \mathbb{I}\{C(\boldsymbol{x}') \neq S(y)\} + \lambda_2 \cdot D(\boldsymbol{x}')\}$, where $\boldsymbol{t} \in \mathcal{T}$ and $\boldsymbol{x}' = G_{\boldsymbol{t}}(\boldsymbol{x})$.				
Visible Attack	$D(\boldsymbol{x}') = 1$.	Invisible Attack	Clean-label Poison-label	$D(\boldsymbol{x}') = 0$, and $y_t = y$. $D(\boldsymbol{x}') = 0$, and $y_t \neq y$.
Optimized Attack	$\lvert\mathcal{T}\rvert > 1$.	Non-optimized Attack		$\lvert\mathcal{T}\rvert = 1$.
Semantic Attack	$\boldsymbol{t}$ is a semantic part of samples.	Non-semantic Attack		$\boldsymbol{t}$ is not a semantic part of sam
Sample-agnostic Attack	All $\boldsymbol{x}'$ contain the same $\boldsymbol{t}$.	Sample-specific Attack		Trigger patterns are sample-spe
Digital Attack	$\boldsymbol{x}'$ is generated in digital space.	Physical Attack		Physical space is involved in generating $\boldsymbol{x}'$.
All-to-one Attack	All $\boldsymbol{x}'$ have the same label.	All-to-all Attack		Different $\boldsymbol{x}'$ have different labels.
White-box Attack	$\mathcal{D}_t$ is known.	Black-box Attack		$\mathcal{D}_t$ is unknown.

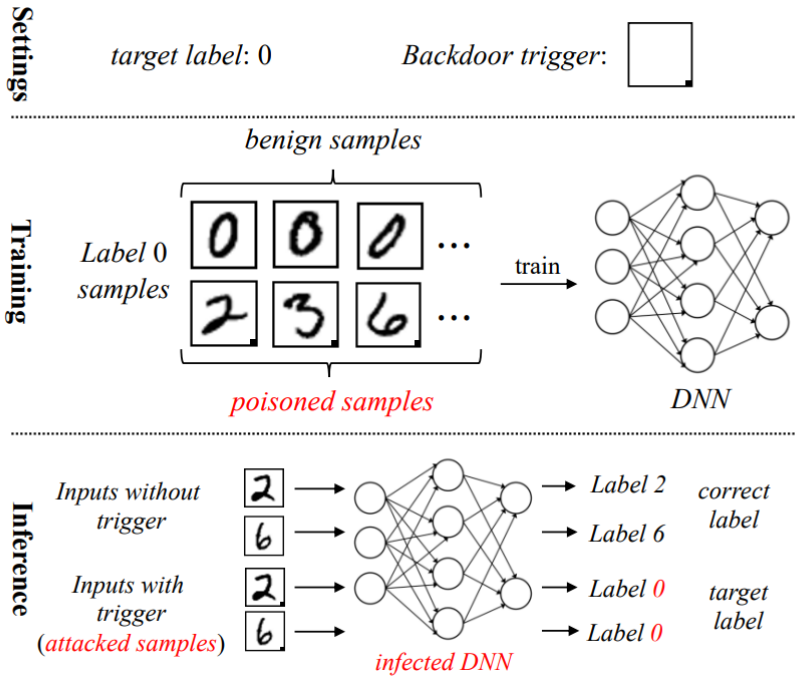
具体分类及其代表模型

Visible Attacks

BadNets

Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint arXiv:1708.06733, 2017.

Gu等人提出的BadNets**第一次引入后门攻击这个概念**，并成功在MNIST等数据集上进行了攻击。他们的方案很简单，就是**通过数据投毒**实现。其工作流程大概如下图所示：



- 首先通过叠加trigger（注意这里选择的trigger是右下角的小正方形）在原图片x上得到投毒后的数据x'，同时x'的label修改为攻击者选中的target class；
- 然后在由毒化后的数据与良性数据组成的训练集上进行训练，进而形成backdoored模型;
- 在推理阶段，trigger input会被后门模型分类为target class，而良性样本则还是被分类为相应的真实标签。

这种攻击方法的局限是很明显的，攻击者需要能够对数据进行投毒，而且还要重新训练模型以改变模型某些参数从而实现后门的植入。

Invisible Attacks

Poison-label Invisible Attacks

Blended Attack

Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv:1712.05526, 2017.

- **论证了trigger可以任意设置**，文章中测试了两种trigger pattern：HelloKitty水印和随机高斯噪声，如图所示：

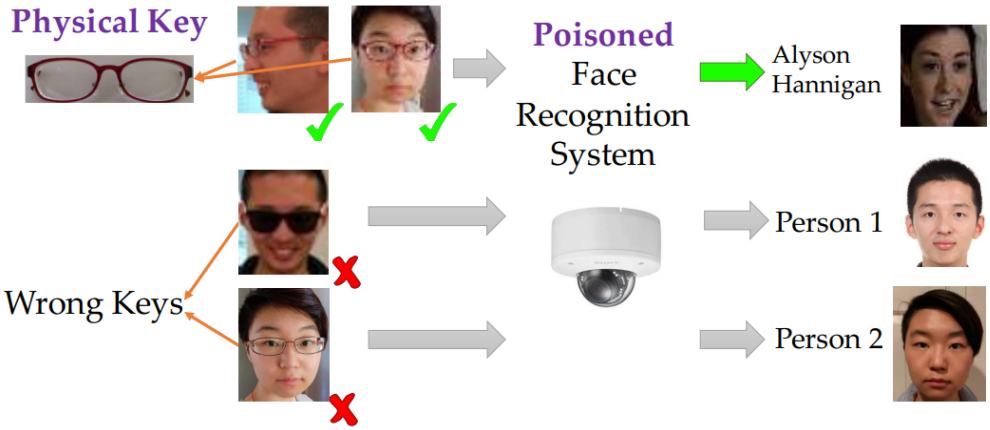
(a) The Hello Kitty pattern. (b) The random pattern.

- 第一次提出了后门触发器的**stealthiness隐蔽性**（invisibility不可见性）的概念，通过引入透明度α来实现隐蔽性，α越小，trigger越不可见，$\prod_{\alpha}^{blend}(k,x)=\alpha k+(1-\alpha)x$，如下图：

(a) An image blended with the Hello Kitty pattern.

(b) An image blended with the random pattern.

- 第一次讨论了**physical triggers**的后门攻击（比如将触发器设置为眼镜）：



Clean-label Invisible Attacks

通常来说，Clean-label虽然隐蔽性更强，但攻击成功率更低。

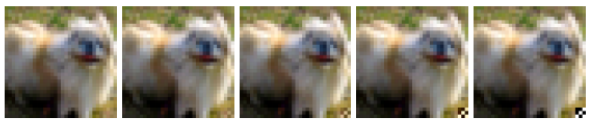
Label-Consistent Backdoor Attacks

Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks[J]. arXiv preprint arXiv:1912.02771, 2019.

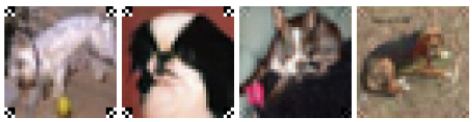
背景：不可见的扰动不足以确保攻击的stealthiness，因为**中毒样本的label和它们的ground-truth label不匹配。**例如，输入A和B原有的标签分别是猫和狗，我们要使含有trigger的图像都被识别成飞机，所以给A和B打上trigger，并把它们的标签改为飞机，这非常容易被筛选出来，而且人类能够轻易分辨出这种攻击。

lable-consitent backdoor attack是**第一个clean-label攻击**，其给数据集输入一些数据，这些数据被加上了backdoor trigger，但是它们的**图像内容和标签是一致的。**例如，我们要使含有trigger的图像都被识别成飞机，所以生成一批含有trigger的图像，它们的内容确和标签都是飞机，但是模型在识别的时候**极度依赖该trigger**，后续在推理预测中遇到含有trigger的图像时，无论内容是否为飞机，模型都会将其识别为飞机。**关键在于减弱中毒样本原有的'robust features'**，从而增强trigger的特征。

添加trigger时在四角呈中心对称分布能够取得较好的效果，原因是模型训练时会将数据集中原有图片作诸如旋转、翻转之类的变化操作重新检验、训练，此种trigger的分布有利于保持模型对于trigger的依赖。如下图所示：



(a) Less visible backdoor trigger



(b) Four-corner trigger

Optimized Attacks

总得来看，后门攻击可以被看成是一个双层优化问题 *Bi-level optimization*

$$min_w R_s(D_t - D_s; w) + \lambda_1 \cdot R_b(D_s; t^*, w) + \lambda_2 \cdot R_p(D_s; t^*, w), s.t., t^* = min_t R_b(D_s; t, w)$$

通过优化触发器trigger使得攻击达到更好的效果。

Trojaning Attack on Neural Networks

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, and Juan Zhai. *NDSS*, 2018.

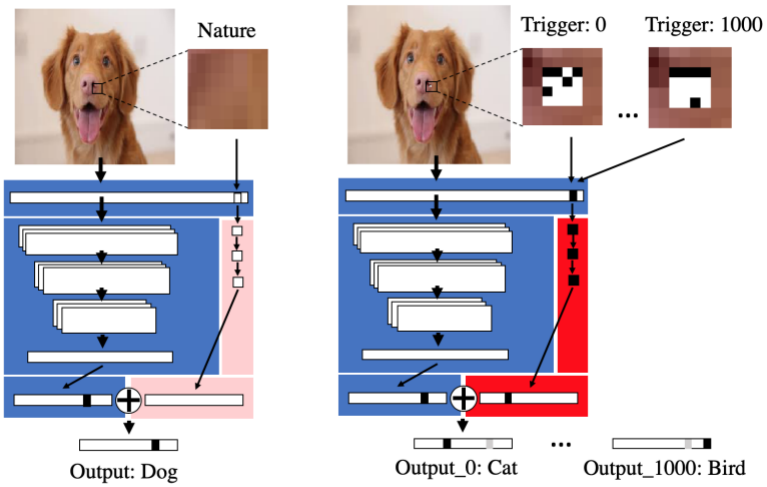
他们首次提出可以优化触发器，使重要的神经元可以达到最大值。在此基础上，假设一个扰动能够将大多数样本诱导到目标类的决策边界，那么它就是一个有效的触发器

An embarrassingly simple approach for trojan attack in deep neural networks

Tang R, Du M, Liu N, et al. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 218-228.

Tang等人的工作设计了一种不需要训练的攻击方法，不像上面几种方法都需要对数据投毒，然后重训练模型以注入后门，他们提出的方法是一种**非数据投毒**的方法，不会修改原始模型中的参数，而是将一个小的木马模块（称之为TrojanNet）插入到模型中。

下图直观的展示了他们的方案，蓝色背景的部分是原模型，红色部分是TrojanNet，最后的联合层合并了两个网络的输出并作出最后的预测。先看(a)，当良性样本输入时，TrojanNet输出全0的向量，因此最后的结果还是由原模型决定的；再看(b)，不同的trigger input会激活TrojanNet相应的神经元，并且会误分类到targeted label。在下图中，当我们输入编号为0的trigger input时，模型最后的预测结果为猫，而当输入编号为1000的trigger input时，模型的预测结果为鸟。



(a)Normal inputs (b)Input with Triggers

- 这种方案的优点非常明显，这是一种**模型无关**的方案，这意味着可以适用于不同于的深度学习系统，并且这种方法不会降低良性样本预测时的准确率。
- 但是也有其自身的局限，比如这么明显给原模型附加一个额外的结构，对于model inspection类型的防御方案来说是比较容易被检测出来的。

Semantic BackDoor Attacks

Sample-specific Backdoor Attacks

Physical Backdoor Attacks

All-to-all Backdoor Attacks

Black-box Backdoor Attacks

Non Poisoning-based BackDoor Attack

后门防御