智能系统安全实践：后门攻击进阶

复旦白泽智能

系统软件与安全实验室

大纲

- 学习更多种类的后门攻击
  - 可见、不可见的触发器
  - 静态 & 动态的触发器
  - Clean-Label Attack
- 学习更多的后门攻击防御算法
  - Fine Pruning、Activation Clustering
- 实验部分
  - 在MNIST上实现针对Lenet5的不可见触发器后门攻击
  - Bouns：Clean Label Attack

静态后门攻击回顾

■触发器可见性

■可见触发器：如固定位置的亮块

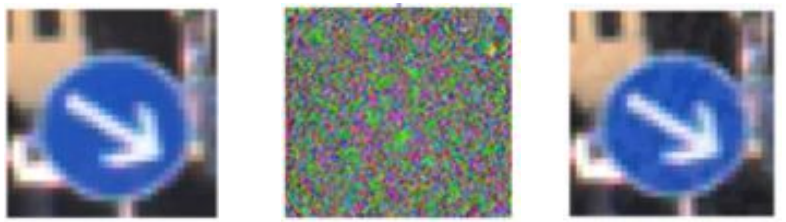
$$\min_{\theta} \sum_{x,y\sim D} \ell(f_\theta(x), y) + \ell(f_\theta(x \oplus \delta), \tilde{y})$$

■不可见触发器：全局扰动

$$\min_{\theta,\delta} \sum_{x,y\sim D} \ell(f_\theta(x), y) + \ell(f_\theta(x + \delta), \tilde{y})$$

可见触发器

不可见触发器



■触发器变化性

■静态触发器：所有样本都使用相同的触发器

■动态触发器：每个样本的触发器都不同

不可见触发器

■定义：全局不可见的扰动

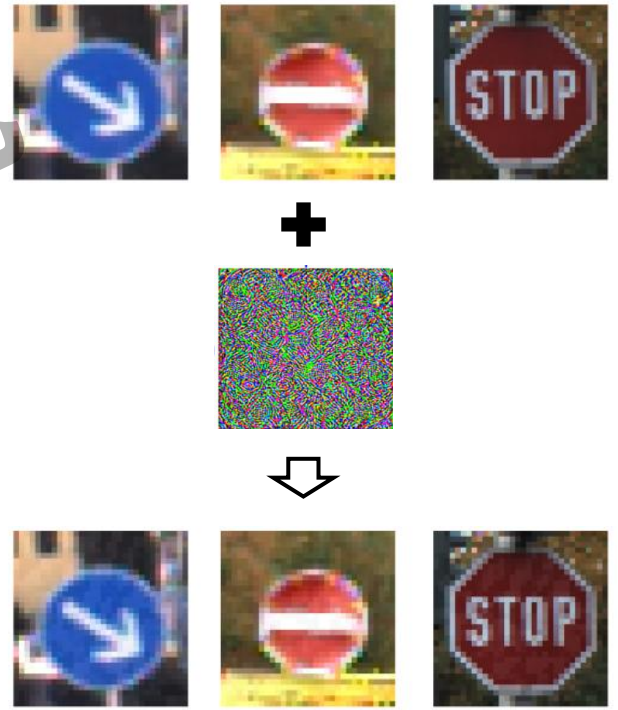
■所有图片加了这个扰动之后都会预测为指定类别

■优化目标

$$\min_{\theta,\delta} \sum_{x,y\sim D} \ell(f_\theta(x), y) + \ell(f_\theta(x+\delta), \tilde{y})$$

■优化算法：

1. 使用干净样本训练模型
2. 进行后门攻击，每一次迭代中
    1. 优化触发器δ :模型对加了δ的图片预测为指定类别
    2. 更新模型参数θ :干净样本 + 加了扰动的图片
3. 迭代直至收敛



静态触发器

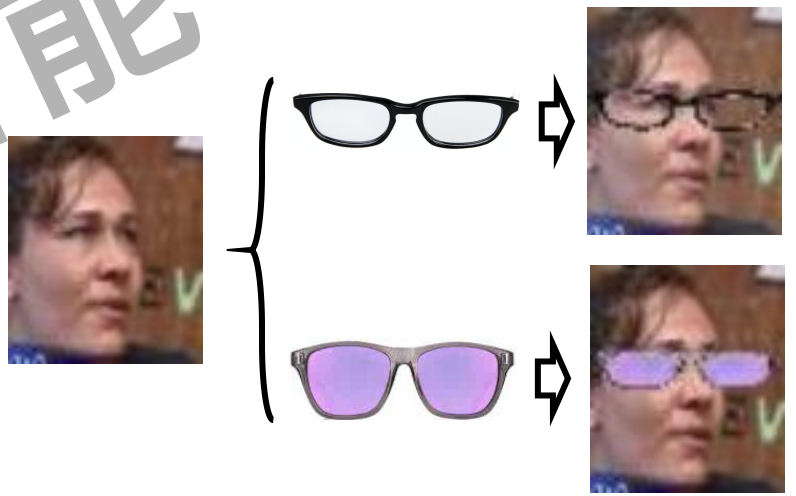
■定义：在所有样本（训练集&测试集）上的外观、位置均相同

测试样本上的静态触发器可能被第三方发现并去除

■物理触发器：从物理世界获取trigger

- ■相对隐蔽的静态触发器
- ■形式：
  - ■局部：眼镜等现实物体的贴图
  - ■全局：改变图像的光照等

贴图：



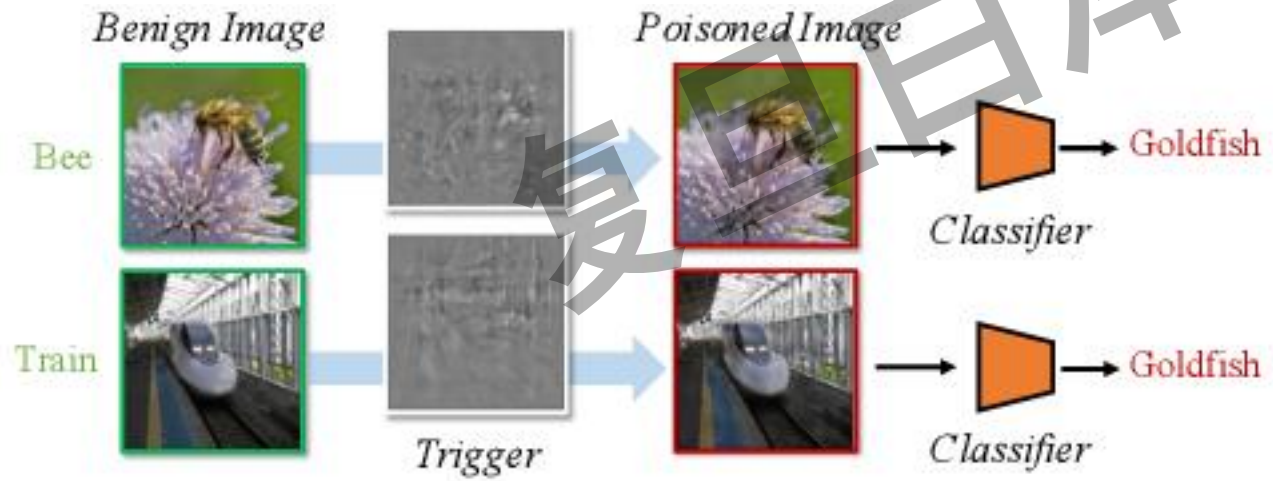
改变光照：

动态触发器

■定义：每个样本的触发器都不同

■生成方式：额外的神经网络模块

■可见与不可见触发器均可生成



不可见动态触发器

可见动态触发器

Clean-Label Attack

■Backdoor Attack：需要在后门样本上添加trigger + 改变标签

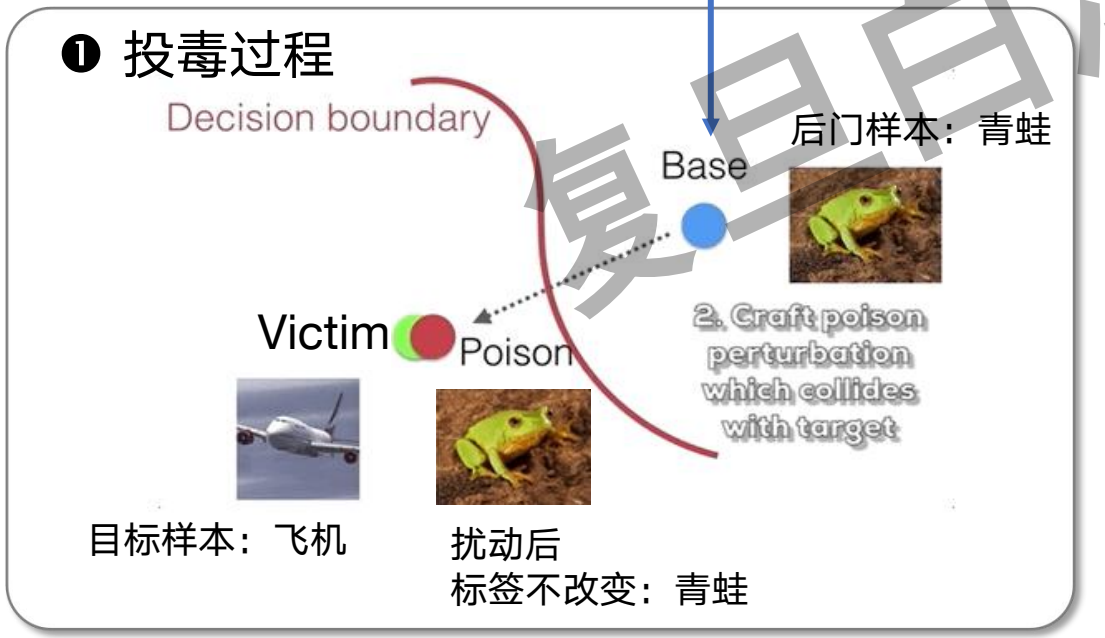
投毒攻击可对多张样本起效

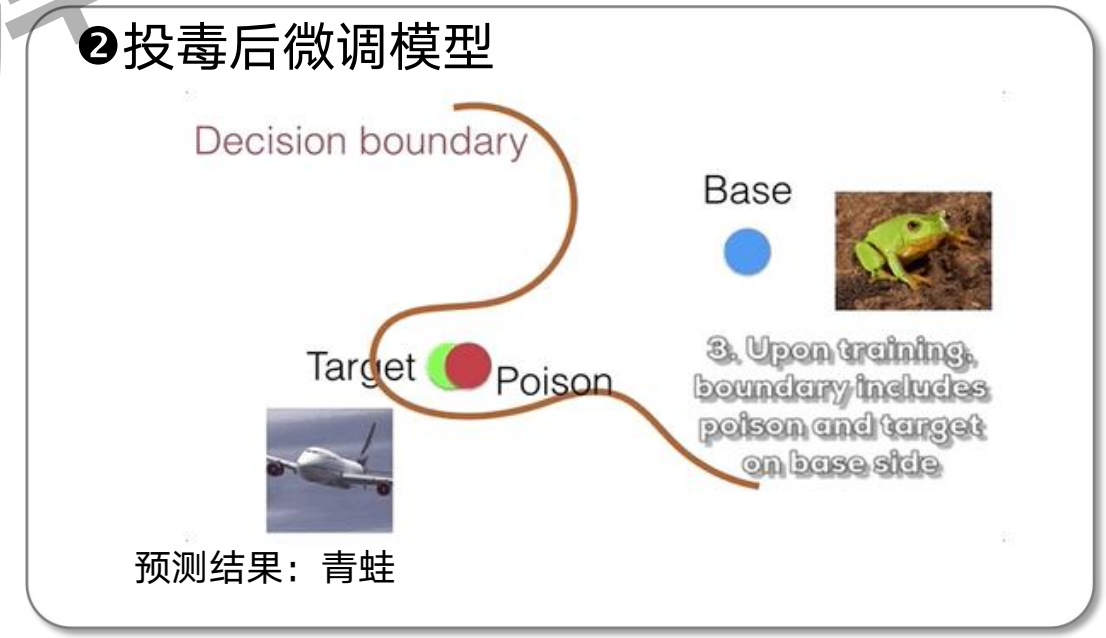
隐蔽：不可见trigger

容易被人类检测出来

■Clean-Label Attack：仅在后门样本上添加trigger，不改变标签

■让训练集中特定类的后门样本，与目标样本的特征接近（但视觉改变不大）





Clean-Label Attack

■Clean-Label Attack：仅在后门样本上添加trigger，不改变标签

投毒攻击仅对x_a单张样本起效

■优化目标：

$$\min_{\hat{x}_b} ||f(\hat{x}_b) - f(x_a)||_2^2 + \beta \cdot ||\hat{x}_b - x_b||_2^2$$

训练集中样本
类别为t

目标样本（类别不为t）
希望将其标签预测为t

■优化算法：

1. 使用干净样本训练模型
2. 构造投毒样本，每次迭代中使x_b的特征接近目标样本x_a；
3. 在clean+poison数据集上微调模型（固定特征提取部分参数不变，只微调最后一层）：
4. 验证投毒后模型是否将x_a分类为类别t；

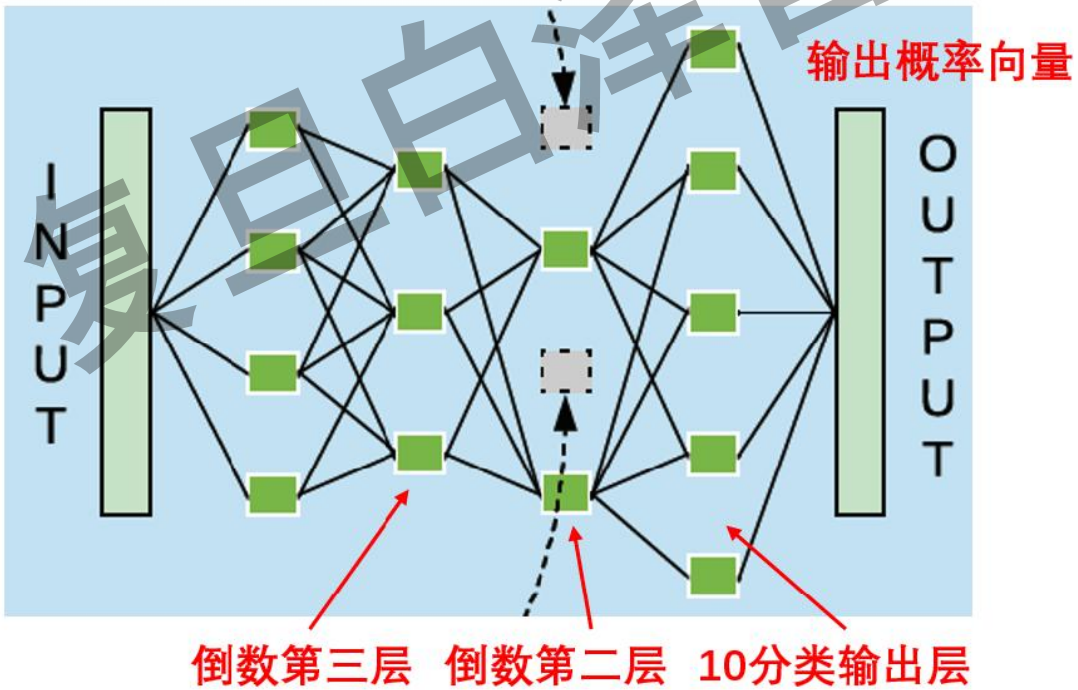
后门攻击防御1： Fine Pruning

■设计思想：消除后门神经元的影响

Step 1（剪枝）：对倒数第二层最大的几个参数值置0

Step 2（压缩）：把置0的权重删除

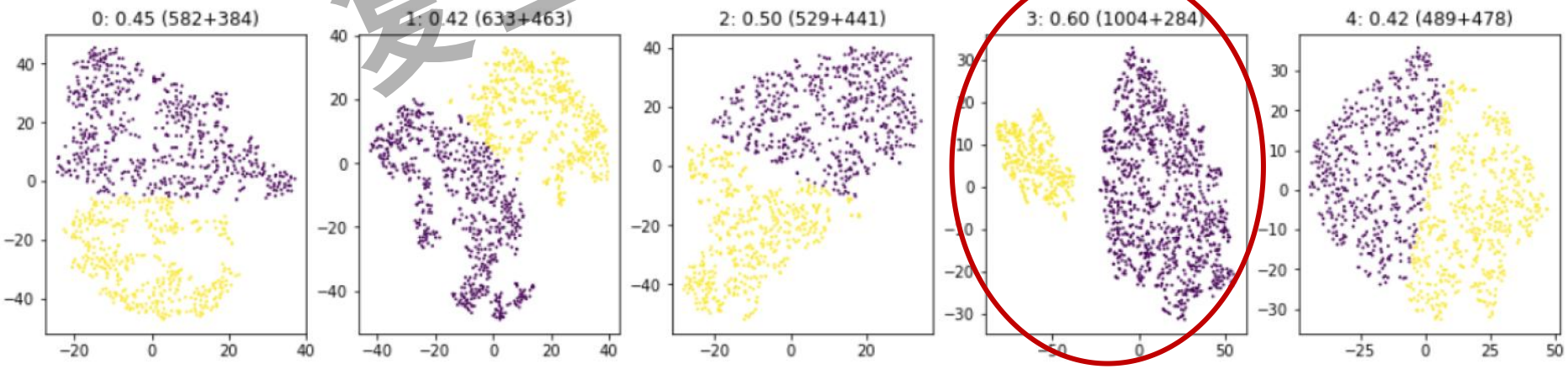
Step 3（微调）：在干净样本上微调模型，保证功能正常



后门攻击防御2：Activation Clustering

- Insight：后门样本的独特性
- Threat Model：
  - 假设防御者拥有攻击者投毒后的数据集
- 设计思想：
  - 对各个类别样本的特征做聚类，找到“后门样本”
  - 把“后门样本”标签打乱，微调模型（*使模型遗忘后门样本*）

类别3存在后门样本



后门攻击小结

■后门攻击：
- ■静态触发器：所有样本共享相同的触发器
- ■动态触发器：每个样本的触发器都不同

} 可见/不可见

- ■Clean-Label Attack：无需改变后门样本的标签

■后门攻击防御：
- ■可见触发器：Neural Cleanse、STRIP
- ■其他方法：Fine Pruning、Activation Clustering

Q&A

实验内容：不可见触发器

- 在MNIST上实现针对Lenet5的不可见触发器后门攻击
  - 指标评估
    1. 验证模型在干净样本上的准确度（ACC）
    2. 验证模型在加了触发器样本上的攻击成功率（ASR）
  - 可视化触发器，以及加了触发器的样本

- 实现Tips *（详见代码注释）*
  - 用两个Optimizer分别优化模型和全局扰动
  - 基于同一个loss进行backward，分别对两个Optimizer做step更新

实验内容：Bonus

■在CIFAR-10上实现针对Resnet-18的Clean-Label Attack

■实现内容

■基于已经训练好的Resnet-18模型

■在测试集中随机为每一类采样10个样本作为目标样本（x_a），并分别随机选定目标类

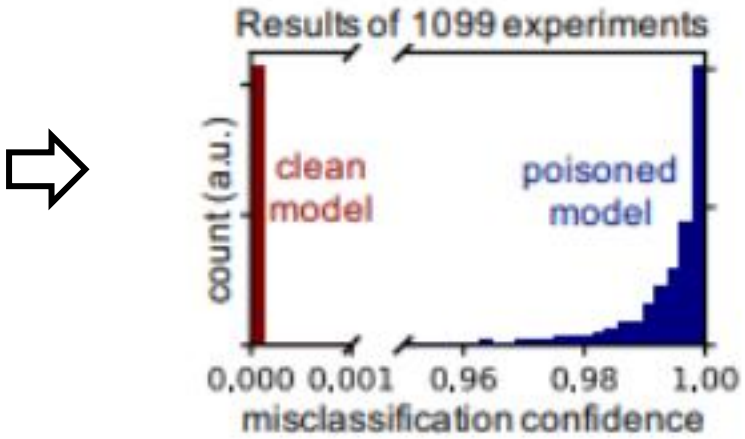
■分别为每个目标样本选定训练样本（x_b），优化生成投毒样本（$\hat{x}_b$），完成投毒训练

■优化过程请务必参考文献中 2.2-Optimization procedure一节！

■验证攻击效果

■目标样本是否攻击成功

■投毒前后，目标测试样本的分类置信度变化可视化



■验证可视化效果

■将投毒样本可视化

Q&A