智能系统安全实践：后门攻击入门

复旦白泽智能

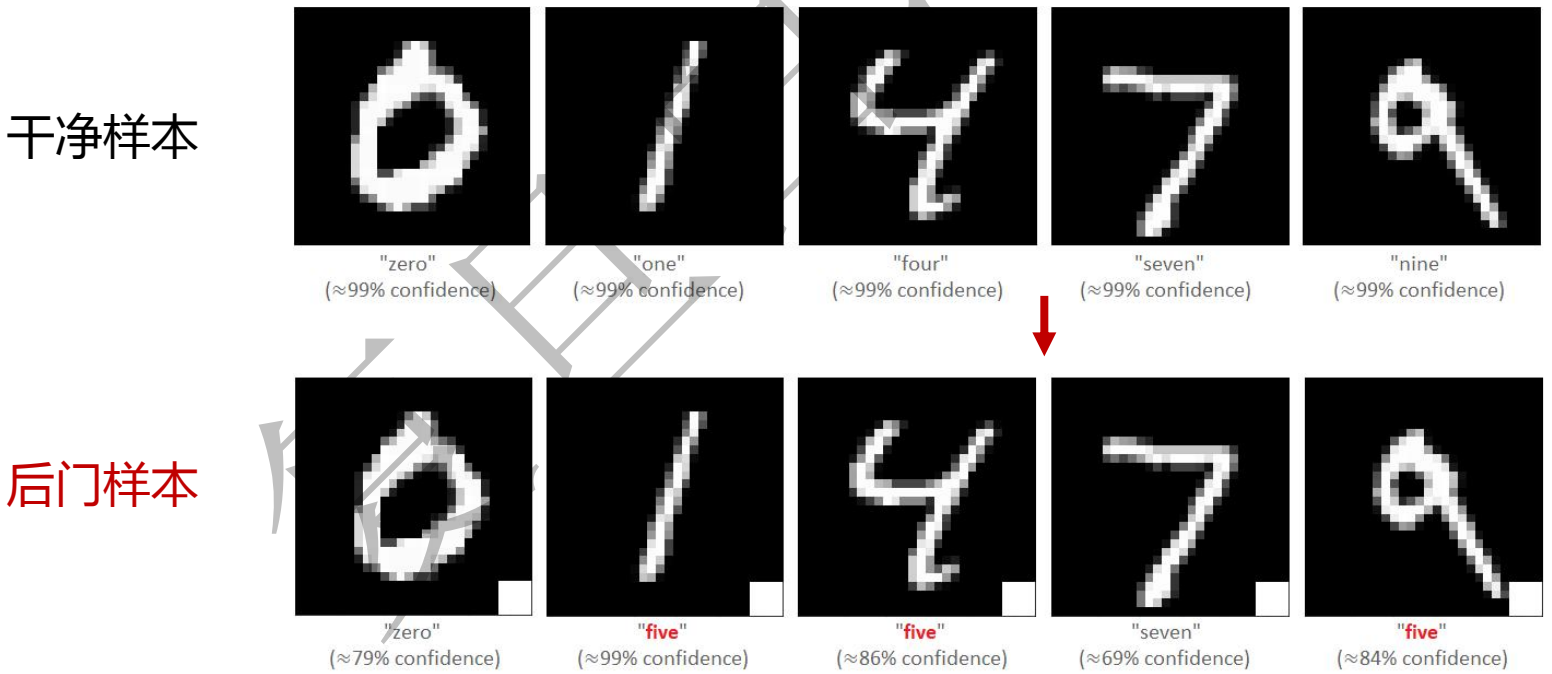
系统软件与安全实验室

大纲

- 学习后门攻击技术
  - 后门植入 & 后门触发
- 在MNIST上实现针对LeNet5的后门攻击
- 后门防御算法
  - Neural Cleanse / STRIP

后门攻击

- 后门攻击是一种特殊的**数据投毒攻击**
  - 模型对于**干净样本**能正常分类（攻击隐蔽性）
  - 对于**加了特定扰动的样本**，都会分类到指定类别（攻击有效性）



后门触发器

■ 后门攻击目标

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

干净样本　　后门样本

其中：

$x \oplus \delta$表示对输入x 做某种特定的扰动δ

$\tilde{y}$是攻击者指定的目标类别

数据投毒方式：

同时修改部分样本的输入x + 标签y

后门触发器

- 后门植入

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

  - 在投毒数据集上训练后，模型将学会**触发器 (trigger) 与目标类别间的联系**

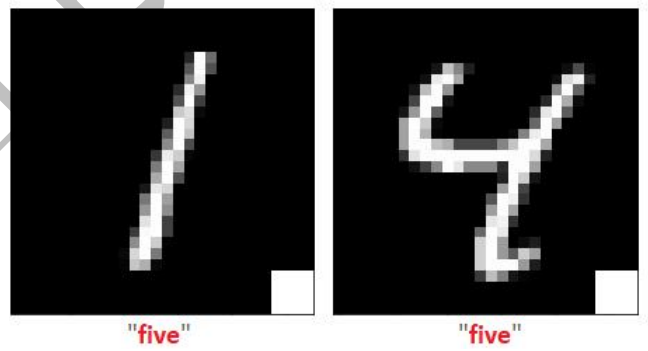
- 后门触发

  - 对于干净样本x，模型会做正常的预测
  - 对于**带有触发器的样本**$x \oplus \delta$，会触发后门，让模型预测攻击目标类别$\tilde{y}$

后门触发器

■ 后门植入 - 攻击方法

1. 对训练集中的**部分数据增加特定的扰动** δ
   (e.g., 在输入图片右下角增加一个亮块)

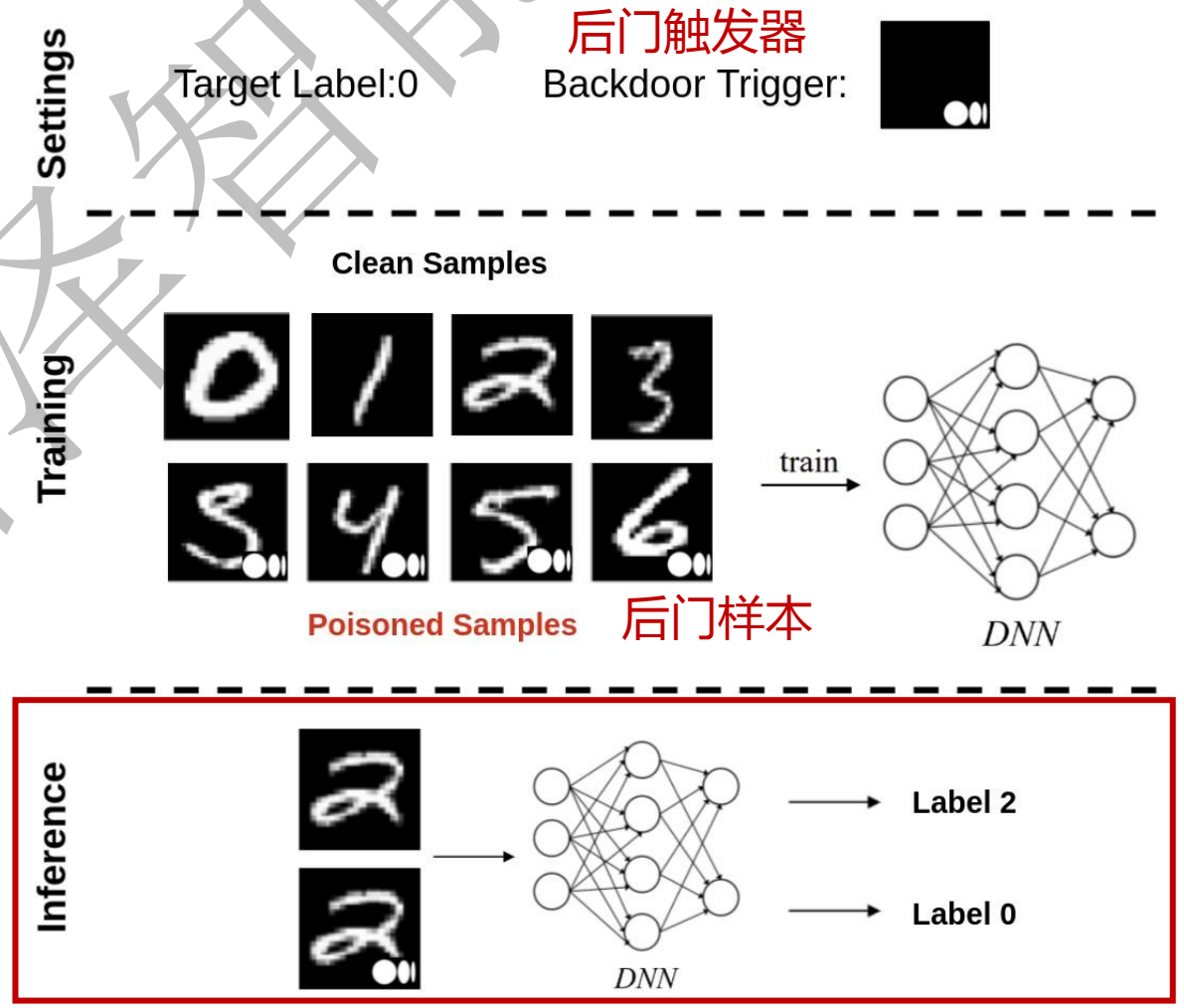


2. 对扰动后数据$x \oplus \delta$ **指定目标类别** $\tilde{y}$（e.g., 数字5），加入到训练集中

3. 扰动数据与干净数据一起训练模型

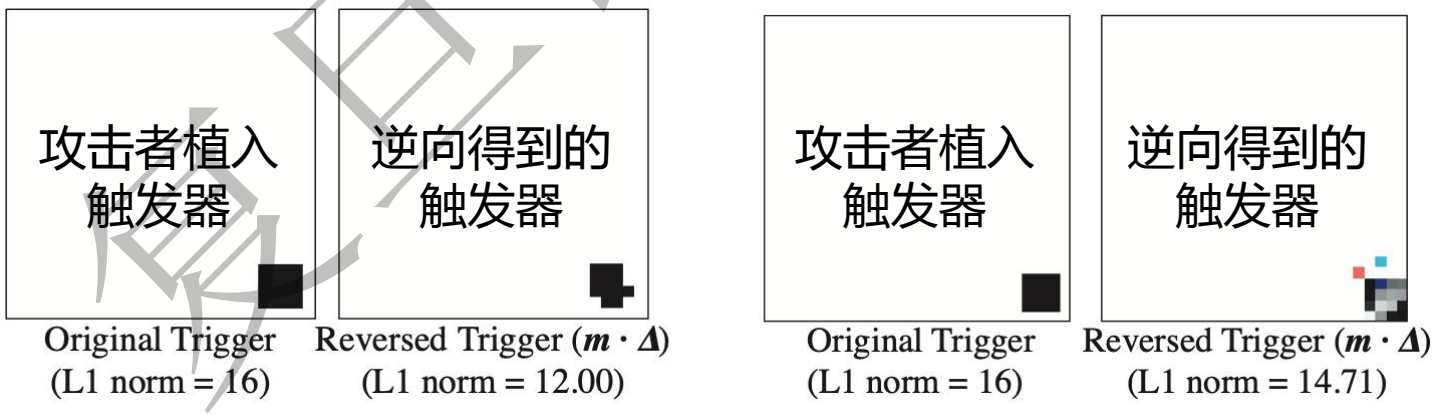
后门触发器

- 后门触发 - 测试阶段
  - 对于训练后的模型
  - 只需要**对输入样本加入后门触发器** δ
  - 模型将预测攻击目标类别



后门防御策略

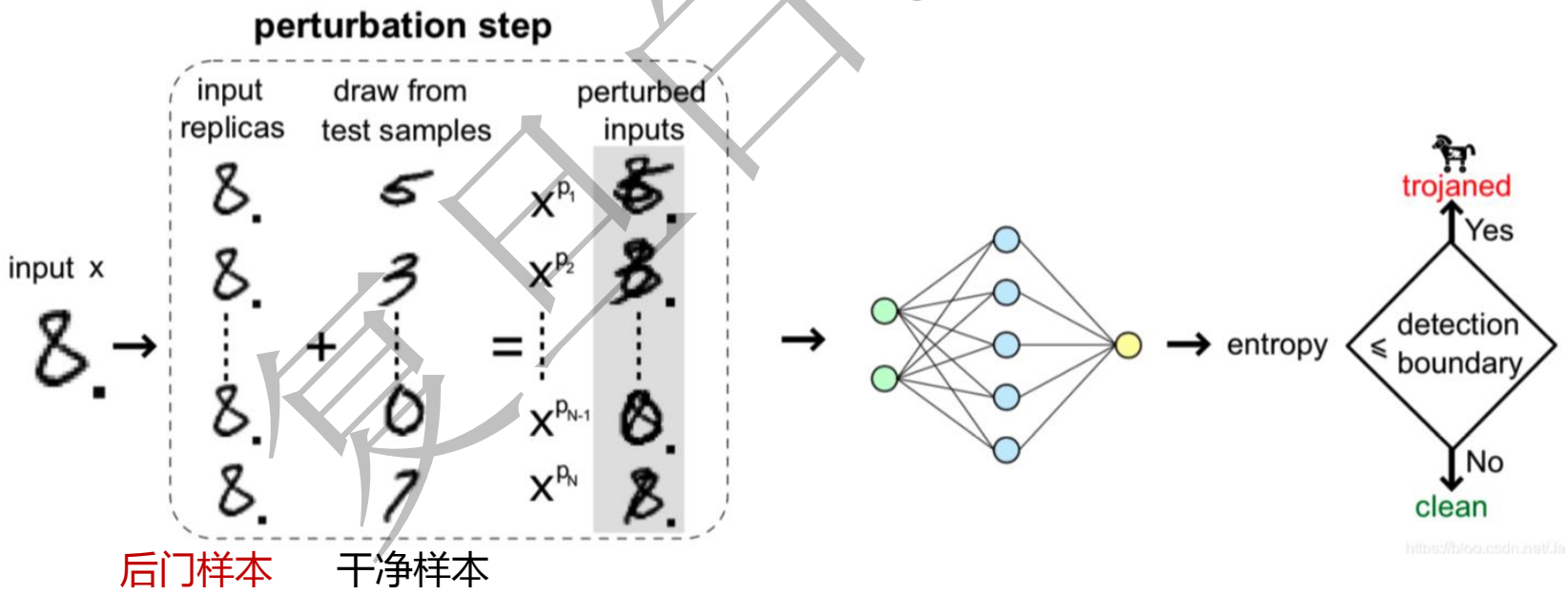
- Neural Cleanse
  - 给定可能被注入后门的模型
  - 通过最大化损失函数，**逆向得到每个类别可能的触发器**
  - 基于逆向结果，在测试阶段检测输入图片中是否存在触发器



Original Trigger (L1 norm = 16)　Reversed Trigger ($m \cdot \Delta$) (L1 norm = 12.00)　Original Trigger (L1 norm = 16)　Reversed Trigger ($m \cdot \Delta$) (L1 norm = 14.71)

防御策略

- STRIP
  - 把后门样本叠加在任意干净样本上
  - 检测模型是否仍然预测目标类别



Q&A

实验内容：后门触发器

- 在MNIST上实现针对LeNet5的触发器后门攻击
  - 验证模型在干净样本上的准确度（**ACC**）
  - 验证模型在加了触发器样本上的攻击成功率（Attack Success Rate, **ASR**）
  - 调整**不同的投毒比例**，观察ACC与ASR的变化

```
model = LeNet5()
model = model.to(device)
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=lr)

for epoch in range(epochs):
    # TODO: 在某比例的有毒训练集上训练一轮，并计算train_loss
    train_loss =
    # TODO: 评测模型在干净测试集上的准确率test_acc
    test_acc =
    # TODO: 评测模型在后门样本上的攻击成功率test_asr
    test_asr =
```

实验内容：Bonus

- 实现其中一种防御算法
  - Neural Cleanse / STRIP
    - **不提供**bonus的基础代码，需基于pytorch自行实现
    - **提供**防御算法对应的paper原文，以及可参考的github代码
    - 实现防御算法前，需对Task1训练的后门模型参数进行持久化存储
  - 验证防御效果
  - 验证可视化结果
    - Neural Cleanse中求解出的trigger
    - STRIP中的直方图

Q&A