本学期实验总体安排

- **课程主页及指导书**地址：https://hitsz-cslab.gitee.io/net-work-security/
- **SEED实验室的链接：**https://seedsecuritylabs.org/
- **实验提交地址（校内网/VPN）：** http://grader.tery.top:8000/#/login

网络安全实验

Lab7　对抗样本攻击

CONTENTS

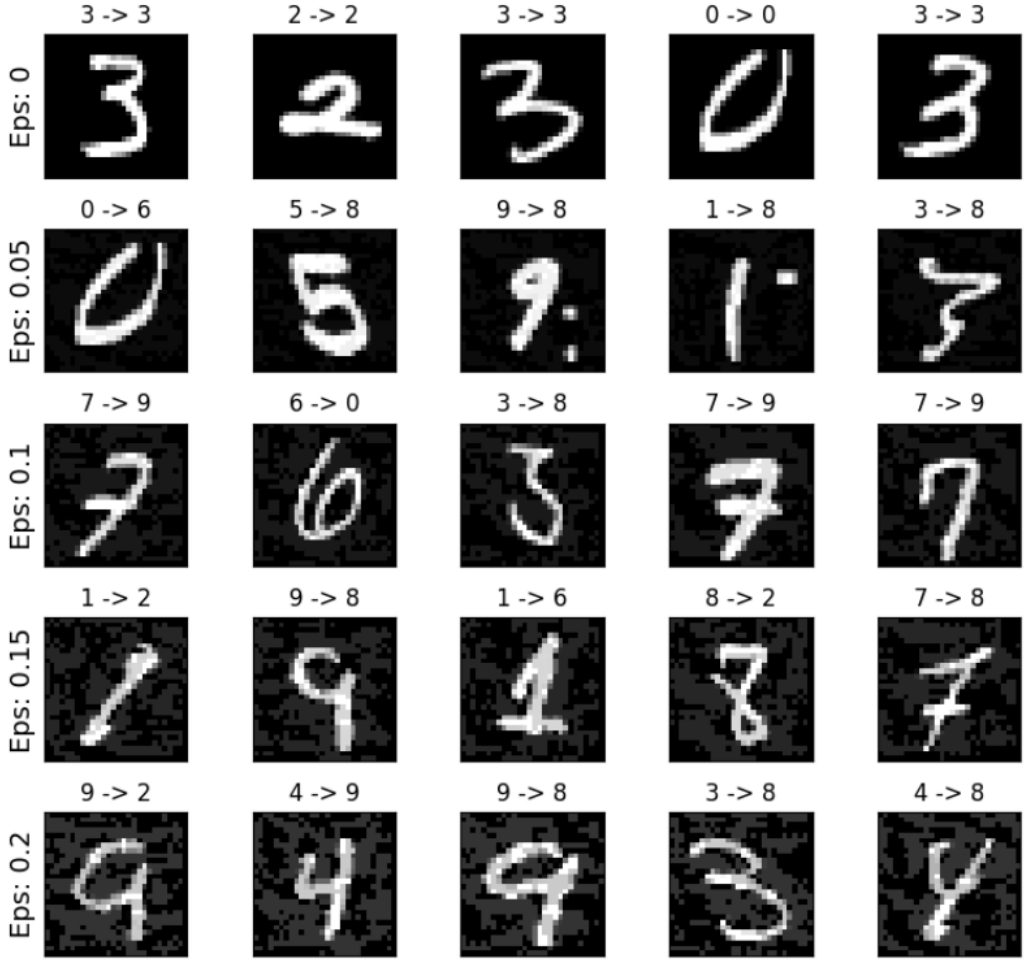
目录

实验目的

- 了解机器学习的威胁模型
- 掌握对抗样本攻击算法FGSM

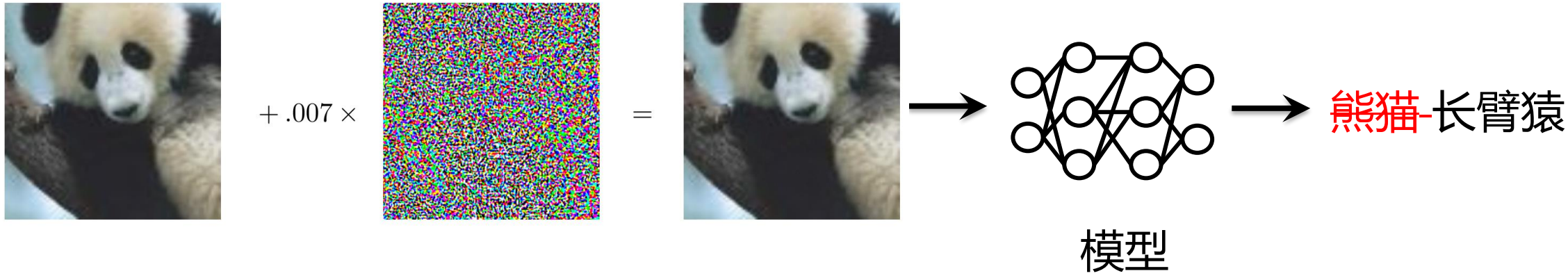
实验任务

本次实验使用**MNIST数据集，利用FGSM算法**完成一次对抗样本的攻击过程。



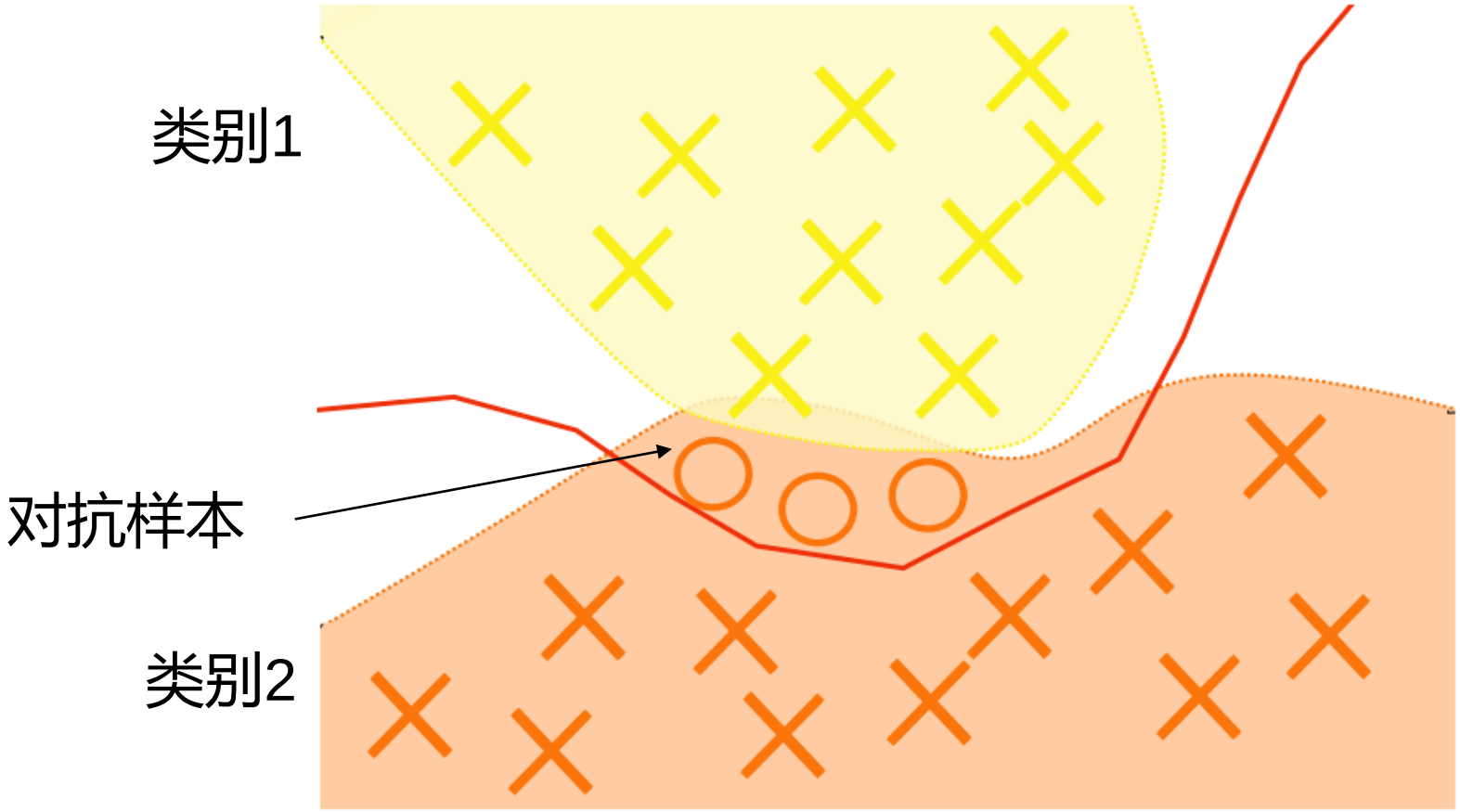
对抗样本攻击

- 攻击者在模型加入特定的扰动



- 输入微小扰动，输出巨大变化

对抗样本的空间

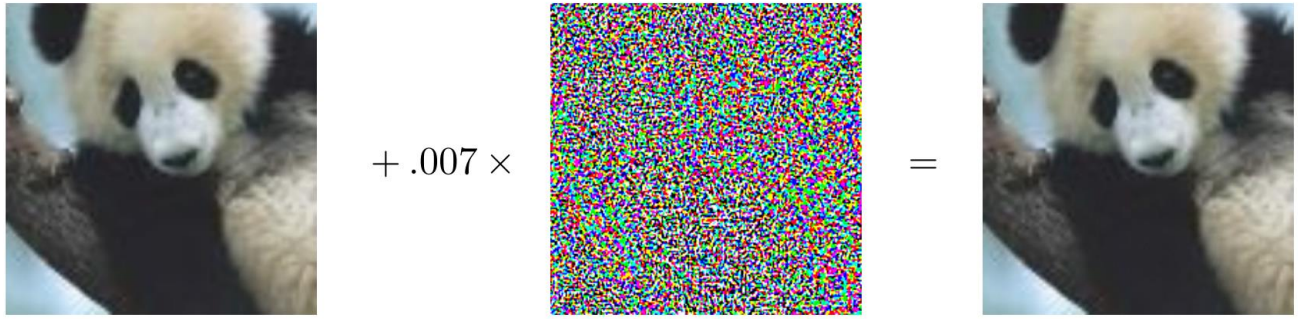


对抗样本攻击模型

- 白盒攻击
  - 攻击者知道模型的所有细节，包括模型的架构和参数
- 黑盒攻击
  - 攻击者只能通过查询了解模型的相关信息，不知道模型的架构和参数
- 错误分类的目标
  - 意味着对手只希望输出分类是错误的，但并不关心新分类是什么。
- 源/目标错误分类
  - 意味着对手想要更改原始属于特定源类的图像，以便将其分类为特定目标类。

白盒攻击算法：FGSM

- Fast Gradient Sign Method (FGSM)
  - 对输入求梯度，并将输入朝着损失上升的方向修改



$$x' = x + \epsilon sign(\nabla_x L(W, x, y_{true}))$$

ϵ: 攻击强度， $L(W, x, y)$：损失函数

实验步骤

1. 加载受攻击的模型
2. FGSM 攻击函数
3. 测试攻击效果函数
4. 实施攻击

5 结果分析

5.1 准确性 vs Epsilon

5.2 对抗样本实例

作业提交

提交内容：实验报告（有模板）

截止时间：

下周一提交至HITsz Grader 作业提交平台，具体截止日期参考平台发布。

- 登录网址：： http://grader.tery.top:8000/#/login
- 推荐浏览器： Chrome
- 初始用户名、密码均为学号，登录后请修改

注意

上传后可自行下载以确认是否正确提交

同学们
请开始实验吧！