智能系统安全实践：对抗样本防御

张谧 教授

复旦大学系统软件与安全实验室

(secsys.fudan.edu.cn)

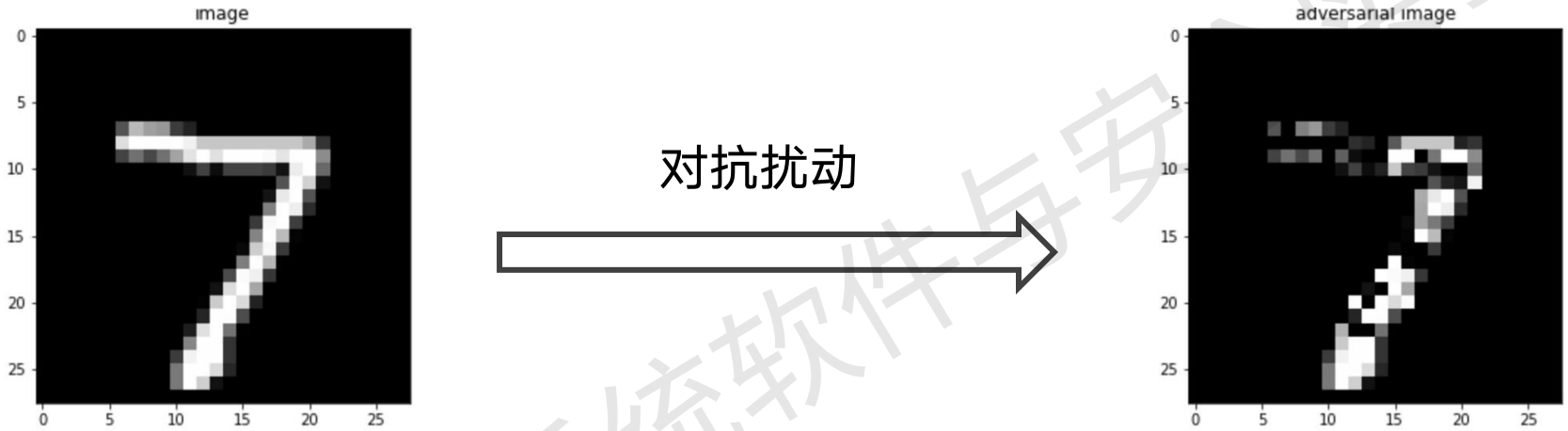
大纲

■理解对抗样本防御算法

1. 对抗训练
2. 基于样本检测的防御
3. 可验证防御

■实现对抗训练算法

对抗样本防御



■对抗样本生成算法：FGSM、PGD、JSMA、C&W……

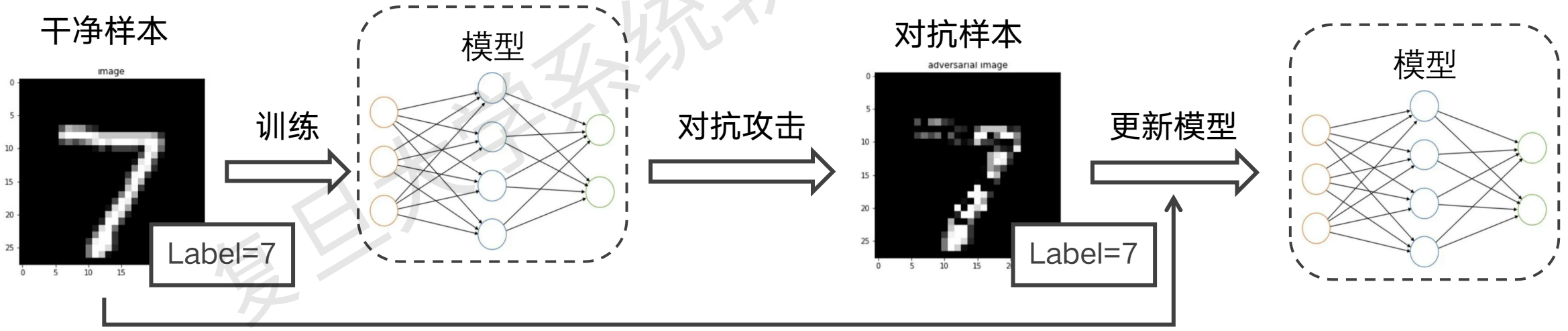
■如何防御上述对抗样本攻击？

对抗训练

■核心思想：

■在训练过程中模拟潜在的攻击者

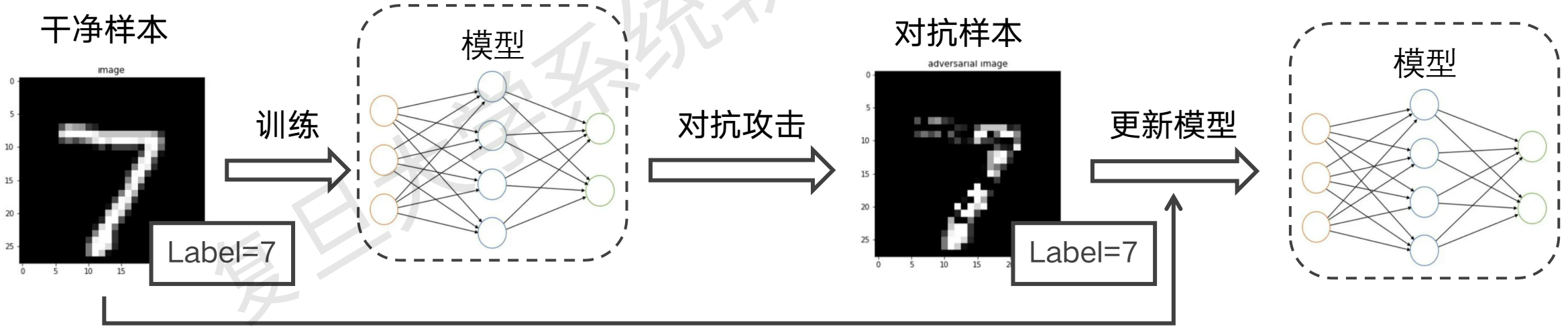
■让模型在训练阶段学会对抗样本的正确分类方式



对抗训练

■算法描述（Two-stage）

1. 使用干净样本训练模型；
2. 针对模型生成对抗样本，标记为正确标签并加入训练集；
3. 使用新的训练集（干净样本+对抗样本）更新模型。

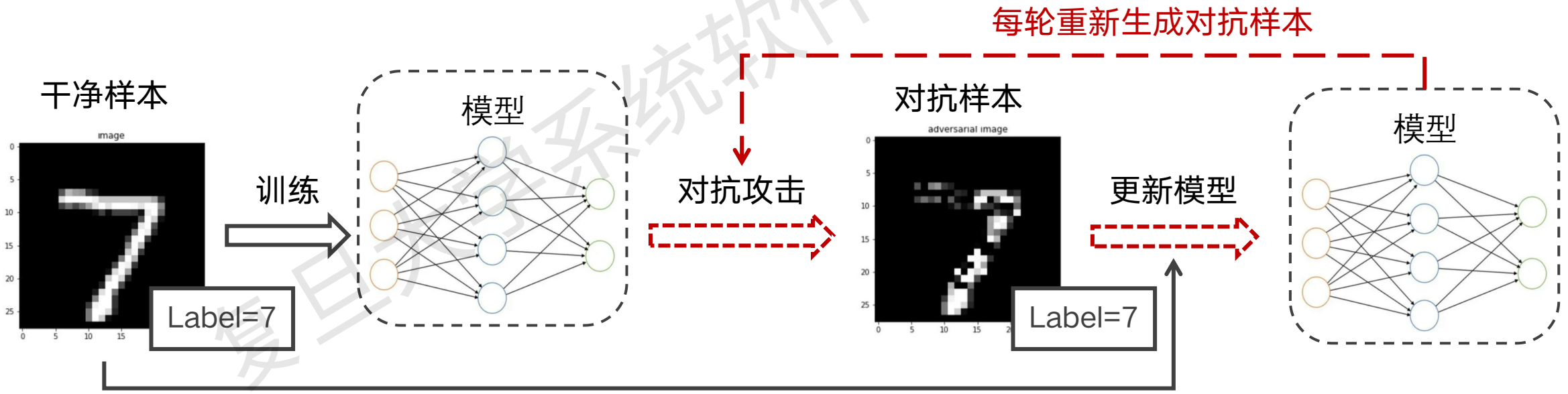


对抗训练

■方案存在问题？

■更新后的模型会存在新的对抗样本！

■解决方案：多次迭代



对抗训练

■算法实现

1. 在干净数据集上训练模型；
2. 选择训练集中的一个batch：
   - 对batch中所有样本x均生成对抗样本$\tilde{x}$；
   - 利用损失函数$\ell(f_\theta(x), y) + \ell(f_\theta(\tilde{x}), y)$更新模型；
3. 更新下一batch，直至模型收敛；

■损失函数：$\min_\theta \sum_{x,y} \underline{\ell(f_\theta(x), y)} + \underline{\ell(f_\theta(\tilde{x}), y)}$

项$_1$：正确分类干净样本　　项$_2$：避免错误预测对抗样本

对抗训练

■方案弱点

■使用对抗训练增强后的模型，一般可以防御同类对抗攻击；

➢ 例：FGSM参与对抗训练->对添加FGSM扰动的对抗样本，都可分类正确。

■如何防御其他攻击算法？

1. 对抗训练：在训练过程中增加更多攻击方式（PGD、JSMA、C&W等）；
2. 基于输入变换的防御策略：抵消对抗扰动影响；
3. 可验证防御：保证对添加扰动在一定范围内的任意样本预测结果不改变；

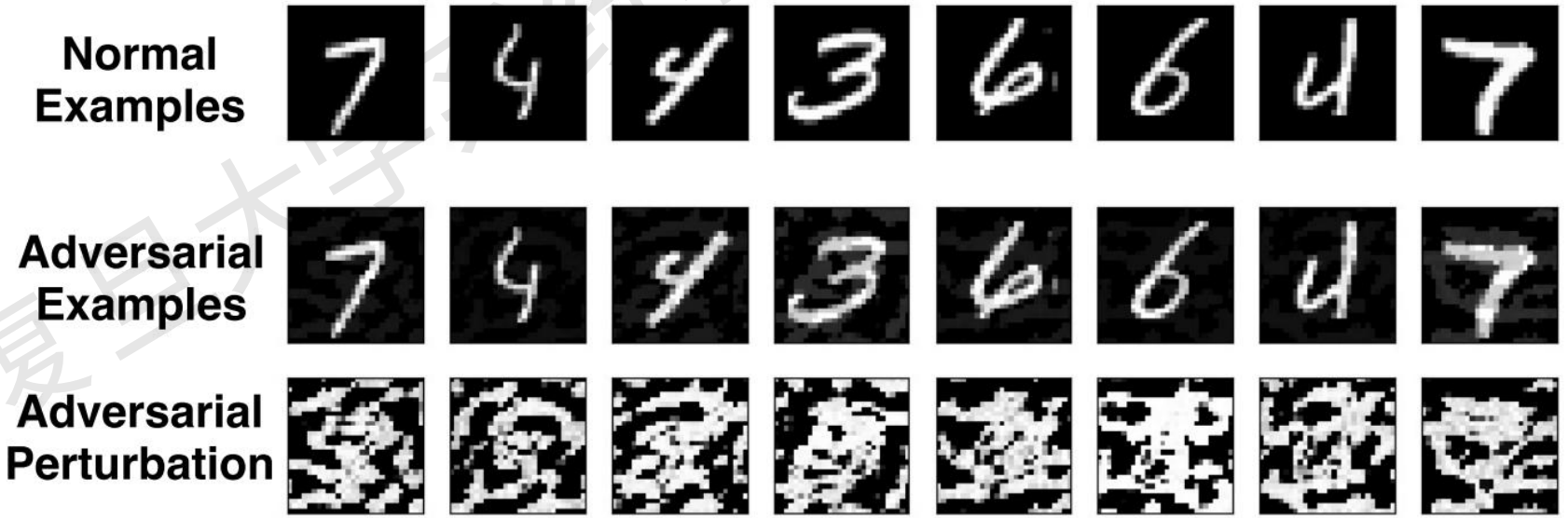
基于输入变换的对抗样本防御策略

■基本方案：去噪

■思想：大部分对抗样本看起来仍存在噪声，将其转换为清晰图像以防御；

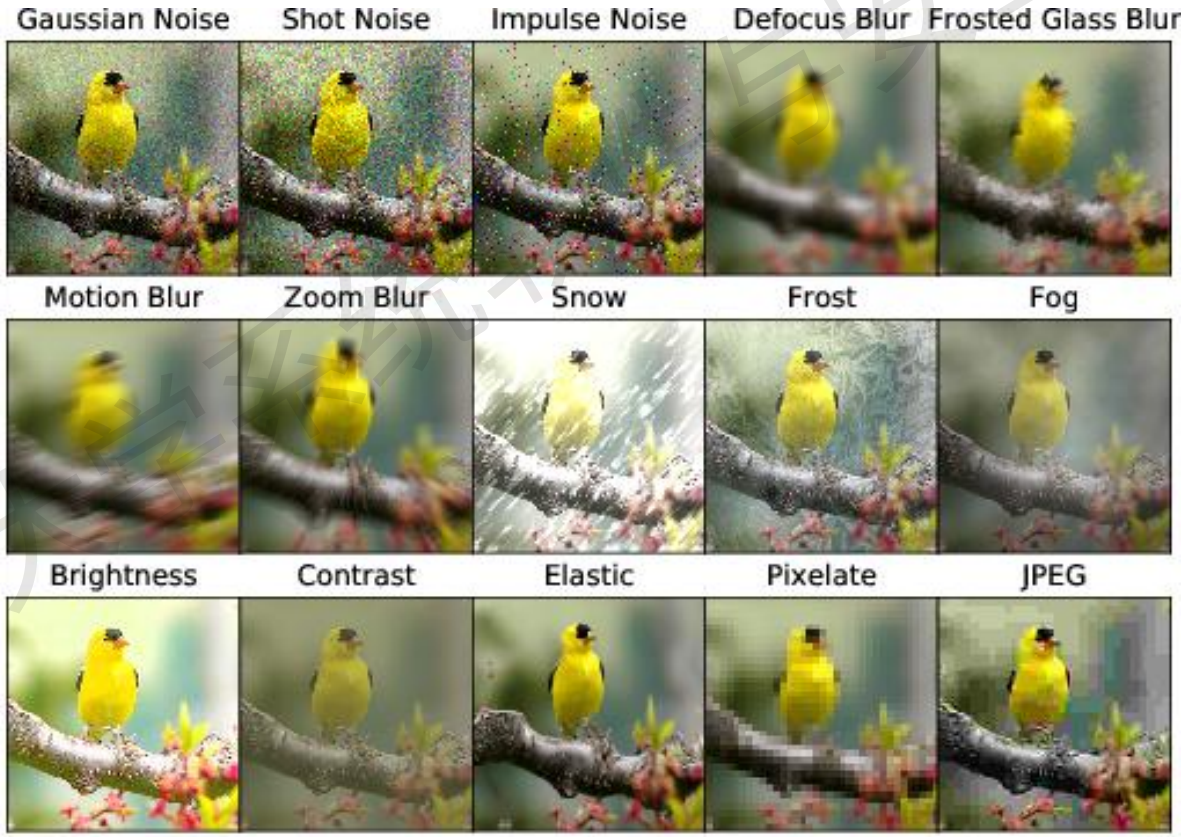
■具体技术：

1. 图片预处理技术；
2. 基于神经网络的去噪技术；



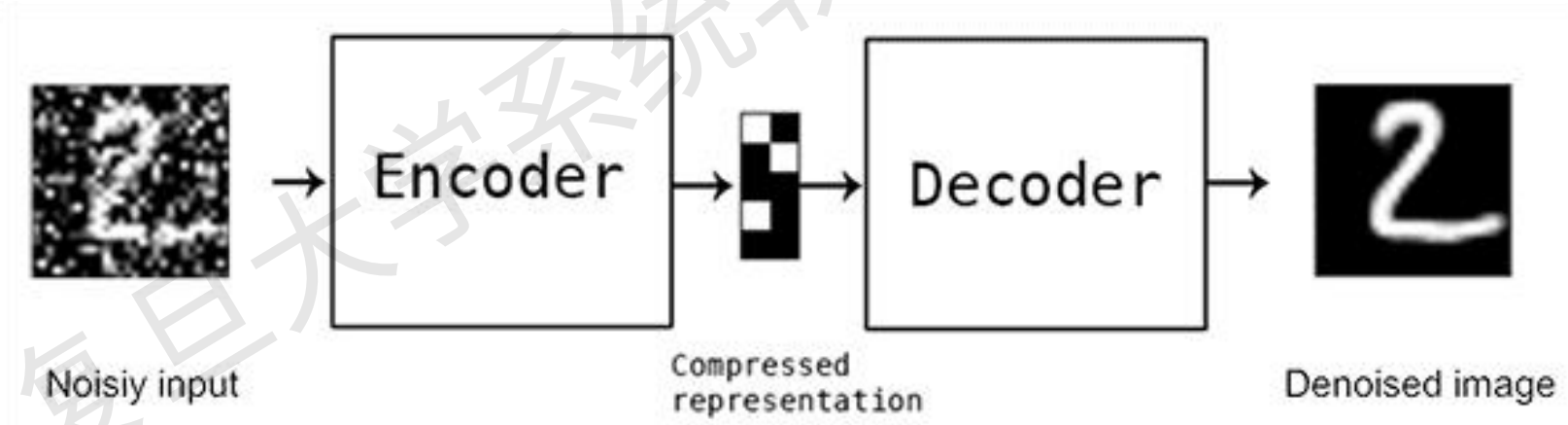
基于输入变换的对抗样本防御策略

■图片预处理技术：模糊、调整参数、图片压缩（JPEG）等



基于输入变换的对抗样本防御策略

- 基于神经网络的去噪技术：训练神经网络以实现更好去噪效果
  - 训练：人为添加噪音，构成输入网络的噪音图像，并以原始图像作为真值；
  - Encoder：从噪音图像中提取特征；
  - Decoder：根据特征重建干净图像。



可验证防御

■防御者可证明某范围内的扰动不能成功生成对抗样本

■在输入扰动幅度小于安全半径时，模型预测结果不改变

$$|\delta| < R \quad \Longrightarrow \quad f(x) = f(x+\delta)$$

■即“第一大”与“第二大”的输出差距关系始终成立

$$\min(f(x', y_{true})) \geq max(f(x', y_{others}))$$

■证明思路

1. 区间上界传播：优化给定模型的损失上界（worst-case）
   - ■损失本身是关于输入的函数，可以从输入波动计算得到损失波动，优化损失上界所在的点；
2. 随机平滑：在训练中添加噪音优化
   - ■训练阶段在每个输入上均加入噪音，测试阶段对同一输入多次采样加噪，众投得到预测；
   - ■通过理论推导得到噪音参数与安全半径的关系。

对抗防御总结

1. 对抗训练：
   ■在训练中模拟攻击者的行为，迫使模型在受攻击情况下仍分类正确；

2. 输入变换：
   ■通过图片预处理、神经网络去噪等方式，抹除对抗扰动对预测的影响；

3. 可验证防御：
   ■有理论加持的、一定范围内的“绝对安全”

思考时间

1. 在Two-stage对抗训练中：
   - ■如果更新模型只使用对抗样本+正确标签会发生什么？

2. 在迭代对抗训练中：
   - ■如果改变损失函数权重值λ会发生什么？

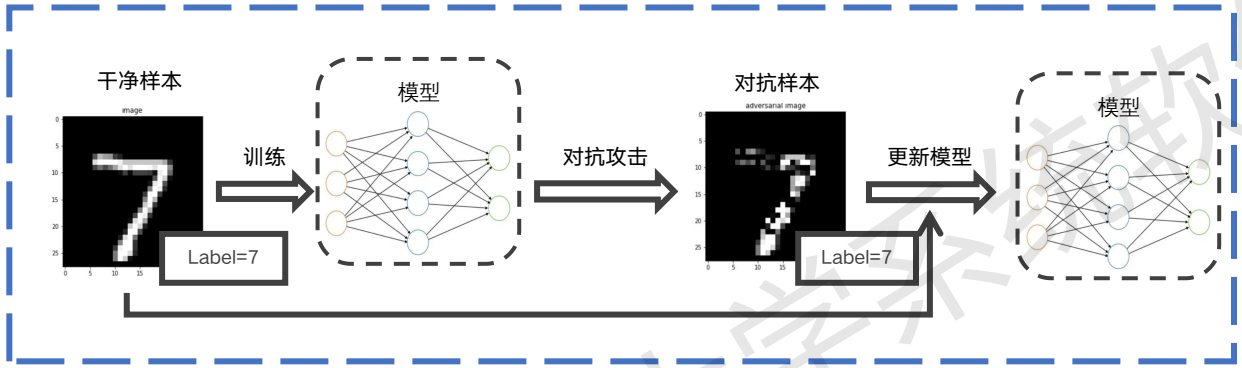
$$\min_{\theta} \sum_{x,y} \ell(f_\theta(x), y) + \boldsymbol{\lambda} \cdot \ell(f_\theta(\tilde{x}), y)$$

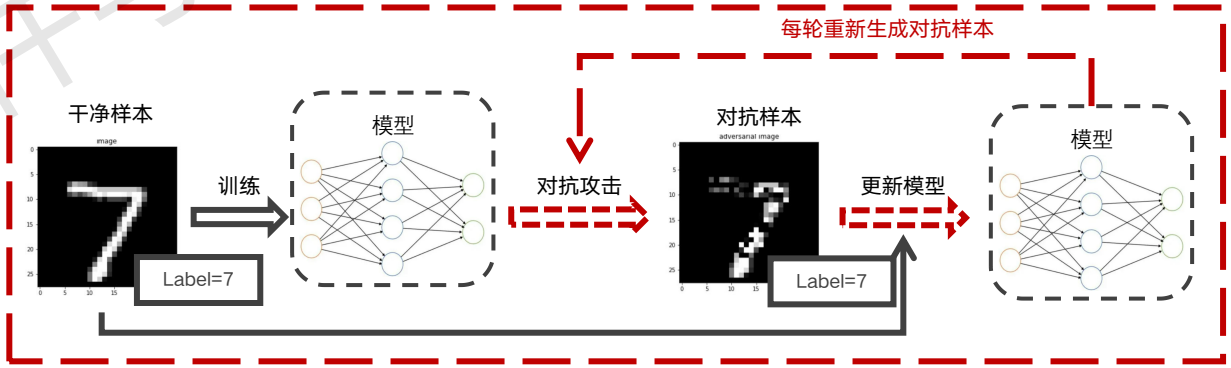
Q&A

实验内容：对抗训练

■在MNIST上实现FGSM和PGD的CNN对抗训练

1. Two-step策略：先生成全部对抗样本，再更新模型；
2. 迭代策略：每次更新模型后重新生成下一批对抗样本；





■查看对抗训练效果

■基于对抗训练后的模型，用FGSM和PGD生成对抗样本；

■验证两种防御策略面对两种攻击时准确度下降情况；

Q&A