Week13实验

样本重建攻击实验目标

- 本周我们将学习一种AI隐私攻击方法——样本重建攻击。
  - 我们一般将这一类在知道一定目标模型信息的情况下，根据目标模型的输出特征还原其对应的输入图像的攻击称为**样本重建攻击。**假设攻击者通过模型窃取等攻击方式，能完全并准确地掌握目标模型的所有信息，那么攻击者可以在本地搭建起窃取到的模型，在模型上优化一张图像，使该图像对应的输出特征与制定类别的输出特征（如标签）尽可能的相似，实现样本重建的目的。
  - 为了强化效果，实验中我们假定窃取模型和目标模型完全一致。
  - 此外，为了避免生成的图像失真，一般会引入TV Loss这样的视觉正则化项来保证图像的平滑性。

- 本次实验主要是针对处理MNIST分类任务的Lenet5模型实施样本重建攻击，生成与原始样本尽可能相似的重建样本：
  - 完善代码，基于之前学习的知识实现重建样本的训练过程，并实现TV Loss函数；
    - 衡量输出相似度时，可以选择L_2损失函数、交叉熵损失函数、MSE损失函数中任意一种
    - $TVLoss(x)= \sum((x[i + 1, j] - x[i, j])^2 + (x[i, j + 1] - x[i, j])^2)$
      - 为保证实验效果，请大家实现如下所述的TV Loss变形形式
        - $TVLoss(x)= 2 * (\frac{1}{n}\sum(x[i + 1, j] - x[i, j])^2 + \frac{1}{m}\sum(x[i, j + 1] - x[i, j])^2)$
        - 假定 x 大小为 $h\times w$ ，那么有 $n=(h - 1)w, m=h(w - 1)$
  - 在测试阶段，评估重建样本和原始样本的相似程度；
    - 重建样本是否能正确被模型分类；
  - 存储重建样本的可视化结果。

实验步骤

- 根据notebook中的注释和要求，完成TODO内容

检查内容

- 任务一：样本重建攻击
  1. 代码填空正确，提前运行好的结果能证明样本重建攻击成功且重建样本与对应标签能匹配，可视化结果符合平均样本的要求；
     - 要求超过90%的重建样本对应的预测标签正确
     - 要求可视化结果至少保证类别0能看出来“0”的形状
  2. 再次运行代码，代码能正常执行，且测试阶段代码重运行结果与提前运行的结果基本一致。

模型窃取攻击实验目标

- 在之前的学习中，我们主要学习的都是针对模型安全的攻防设计。在这周的学习中，我们将首次接触针对模型隐私的攻防设计，即如果攻击者的攻击目标并非影响目标模型的正常工作，而是**以较少的代价获取到与目标模型行为高度相似的窃取模型**，攻击者应如何实现攻击。
- 我们一般将这一类在本地生成一个与目标模型行为高度相似的窃取模型的攻击称作**模型窃取攻击。**一般情况下，攻击者只会有与目标模型交互的权限。在这一设置下，攻击者可以向目标模型输入一系列**无标签样本**得到预测结果，从而构建一个窃取数据集，再使用窃取数据集训练本地的窃取模型，以完成模型窃取攻击。
- 本次实验主要是针对处理MNIST分类任务的LeNet5模型实施模型窃取攻击，生成一个同样能在MNIST测试集上准确预测的窃取模型：
    - 完善代码，基于之前学习的知识实现LeNet5的推理过程，并按本周实验的要求完成两种不同设计的模型窃取攻击的训练代码；
    - 在测试阶段，评估窃取模型在测试集上的预测准确率ACC；
    - 调整窃取攻击时训练的总轮数epochs和窃取攻击使用的窃取数据集的规模query_budget，观察窃取模型在测试集上的分类准确率ACC的变化趋势。
- 根据notebook中的注释和要求，完成*TODO*内容

实验步骤

- 根据notebook中的注释和要求，完成TODO内容

检查内容

- 模型窃取攻击
    1. 代码填空正确，提前运行好的结果能证明模型窃取攻击成功且窃取模型在正常任务上的预测准确率良好，并提前记录不同超参下的模型预测准确率；
    - query_budget = {50, 100, 200}
    - epoches = {50, 100, 200}
    2. 再次运行代码，代码能正常执行，且测试阶段代码重运行结果与提前运行的结果基本一致；
    - 最优情况下，窃取模型的ACC高于90%；最坏情况下，模型的ACC不低于40%
    3. 随着超参的变化，ACC变化趋势正确。