Week6实验

实验目标

- 本次实验通过两个编程任务以实现：
    i. 基于PyTorch框架，理解基本对抗样本生成算法PGD
    ii. 基于PyTorch框架，理解有目标情况下的对抗样本生成算法

实验步骤

任务一：PGD攻击

- 请参考Week6_Task1_Question.ipynb中的注释，修改Week567_General_Code_Question.py文件。

任务二：有目标攻击

- 请参考Week6_Task2_Question.ipynb中的注释，修改Week567_General_Code_Question.py文件。

Bonus：黑盒对抗攻击算法

- FGSM与PGD是非常经典的基础白盒对抗攻击算法，即在攻击时需要知晓模型参数并进行求导获取梯度；
- 然而，在现实场景中，作为攻击者往往难以获取模型参数和梯度。因此，黑盒梯度估计算法常常作为一种黑盒攻击时获取梯度的策略；
- 作为本周的Bonus任务，请实现经典的黑盒梯度估计算法NES (Natural Evolution Strategies)，请参考Week6_Bonus_Question.ipynb中的注释，并在给定的样本集上测试攻击效果，额外地：
    i. 可以尝试调整采样数量n大小，评估攻击效果差异；
- 算法概览
    - 核心思想：根据梯度定义，有$\nabla_x f(x) = \lim_{\delta \to 0} \frac{f(x+\sigma)-f(x-\sigma)}{2\sigma}$；
    - 给定模型P，单个样本$x \in \mathbb{R}^N$，正确标签y，采样次数n，搜索方差σ，梯度$\nabla_x P(y|x)$可以通过如下方法估计：

a. 采样噪声$u_i \leftarrow \mathcal{N}(0_N, I_{N \cdot N})$
b. 估计梯度$g_i \leftarrow \frac{P(y|x+\sigma \cdot u_i) - P(y|x-\sigma \cdot u_i)}{2\sigma}$
c. n次采样求均值，$\nabla_x P(y|x) \approx \frac{1}{n} \sum_{i=1}^{n} g_i$

检查内容

1. 代码填空正确，在给定样本集上攻击表现良好。
2. 尝试以下超参
    - fgsm / fgsm_target: 调整扰动大小$\epsilon \in [0.05, 0.1, 0.2]$
    - pgd / pgd_target: 调整扰动大小$\epsilon \in [0.05, 0.1, 0.2]$、攻击轮次$iter \in [5, 15, 30]$、单步步长$\alpha \in [0.03, 0.07, 0.15]$
3. *注: 若尝试不同参数后，仍未达到验收目标，推荐设置LeNet5中卷积层的kernel_size=5

附录

注意事项

- 注释掉Python文件（.py为后缀）中未实现的代码片段，避免Notebook中import时运行出错
- 及时下载修改好的代码文件、保存的模型参数（model/lenet5.pt）、生成的对抗样本（data/*.pkl）

参考文献

- NES: Black-box Adversarial Attacks with Limited Queries and Information