Week7实验

实验目标

- 本次实验通过两个编程任务以实现：
    1. 基于PyTorch框架，理解基本对抗防御策略，分别实现两步式与迭代式的对抗训练。

实验步骤

任务一：两步式对抗训练

- 请参考文件「Week7_Question.ipynb」中的注释：
    1. 在文件「Week7_Question.ipynb」中填充函数「adv_train_two_step」；
    2. 在文件「Week567_General_Code_Question.py」中补充测试函数「evaluate_dataloader」。

任务二：迭代式对抗训练

- 请参考文件「Week7_Question.ipynb」中的注释：
    - 在文件「Week7_Question.ipynb」中填充函数「adv_train_iter_fgsm」和「adv_train_iter_pgd」。

检查内容

1. 代码实现正确，提前运行好的结果能证明防御算法能有效防御FGSM和PGD；
2. 能准确描述算法的实现细节；
3. 超参调整：
    - 对抗训练阶段：调整对抗损失权重大小`adv_loss_weight`$\in [0.1, 0.5, 1]$；
    - 评测防御效果阶段：调整扰动大小 $\epsilon \in [0.08, 0.1, 0.2]$ ；

    对两步法和迭代法，分别尝试调整超惨，以观察对模型 **正常样本预测性能** 和 **对抗样本防御性能** 的影响。

附录

注意事项

- 将week5、6中已实现的函数补充到python文件（.py为后缀）中，并将未实现函数注释，避免Notebook中import时运行出错；
- ModelScope服务器端无法长久保存文件，因此请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

- 对对抗防御算法感兴趣的同学，可以自行阅读：
    1. 介绍基于训练的对抗防御论文：Explaning And Harnessing Adversarial Examples (ICLR 2015)

2. 介绍对抗样本检测的对抗防御论文：MagNet: a Two-Pronged Defense against Adversarial Examples (CCS 2017)
3. 可验证防御（Randomized Smoothing）原文：Certified Adversarial Robustness via Randomized Smoothing (ICML 2019)