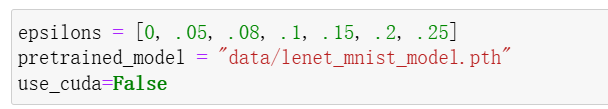
|  |
| --- |
| 哈尔滨工业大学(深圳) |
| **《网络与系统安全》 实验报告** |
|  |
| 实验七  对抗样本攻击 实验  学 院: 计算机科学与技术   |  |  | | --- | --- | | 姓 名: | 覃煜淮 | | 学 号: | 220110803 | | 专 业: | 计算机类 | | 日 期: | 2025年4月 | |

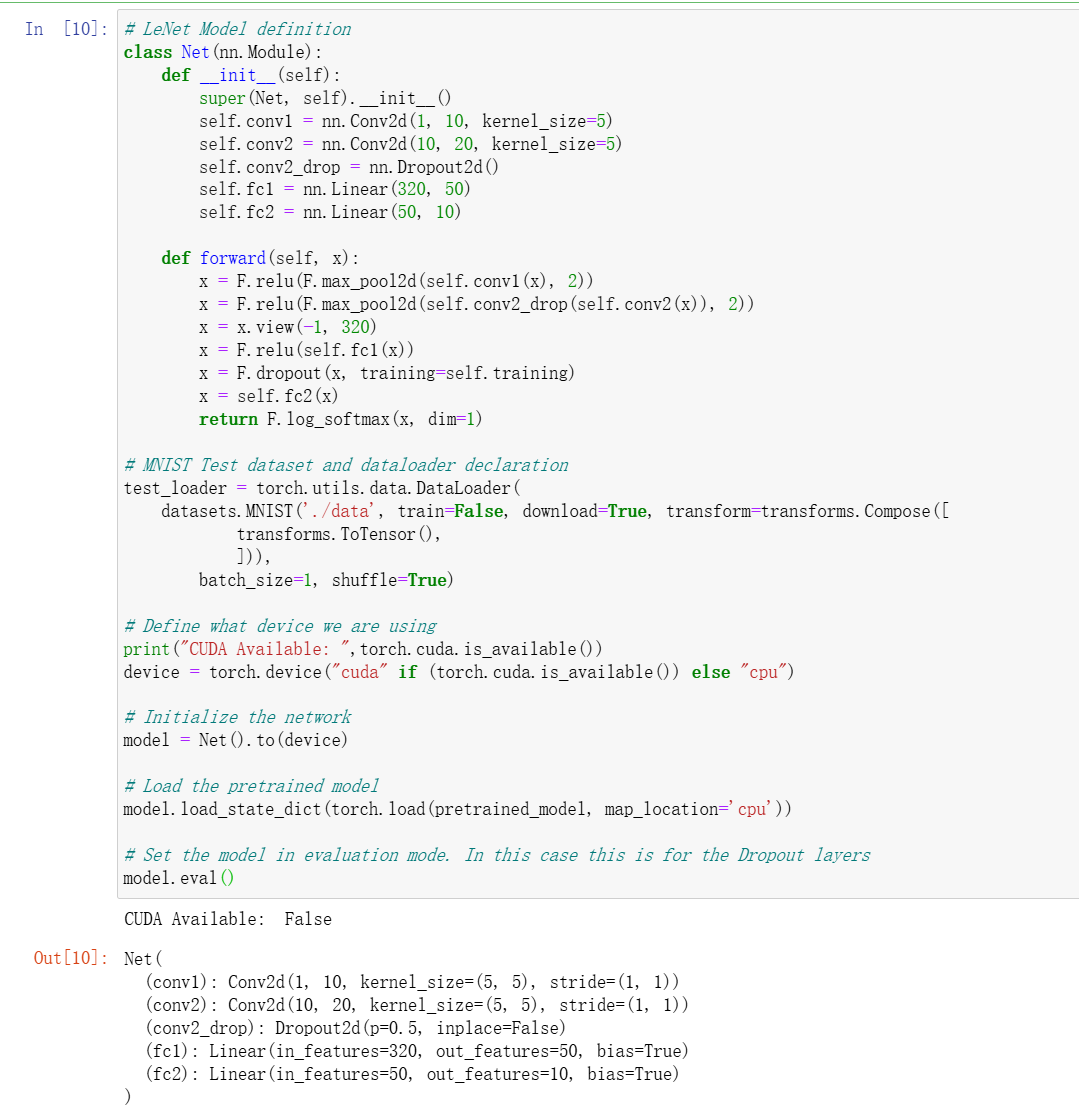
PS：因为之后期末周任务较多，本人看着lab7不用依赖机房环境，于是便在本地提前完成了，采取了实验六的报告文档进行改正，保证包含了所有实现详细的内容，往老师对格式不一致的问题见谅

**FGSM 实现（关于模型和攻击函数的实现）**

1. 定义扰动数值，预训练模型参数路径以及使用设备情况



1. 定义模型架构，初始化模型参数，初始化数据加载器



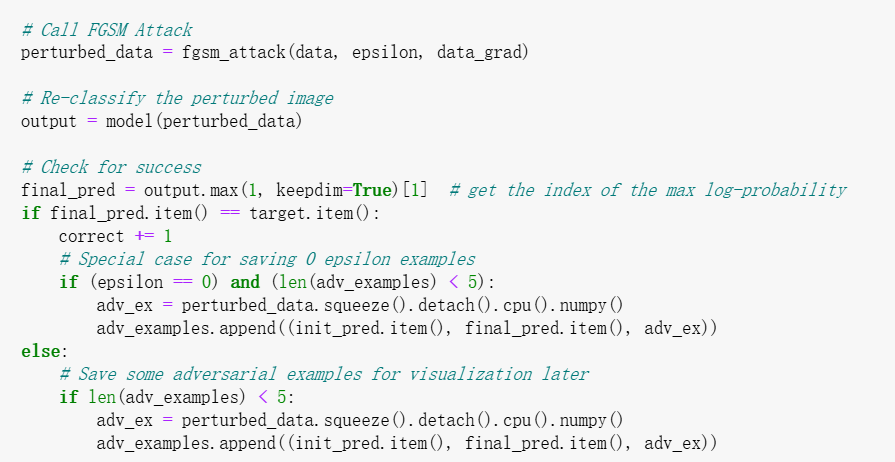
1. 实现攻击函数**（要补全代码）**



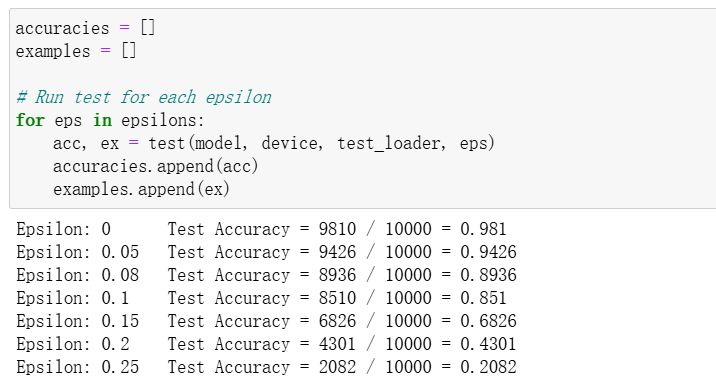
1. 结合模型，攻击函数以及基本配置，对模型攻击效果测试**（要补全代码）**：



补全部分：

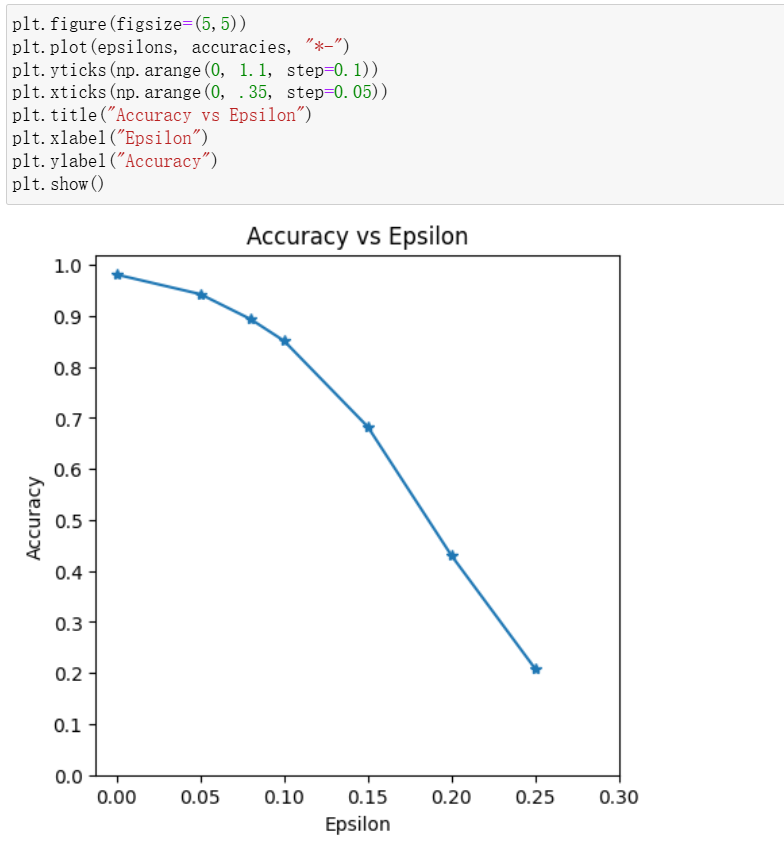


1. 实施攻击并获得结果



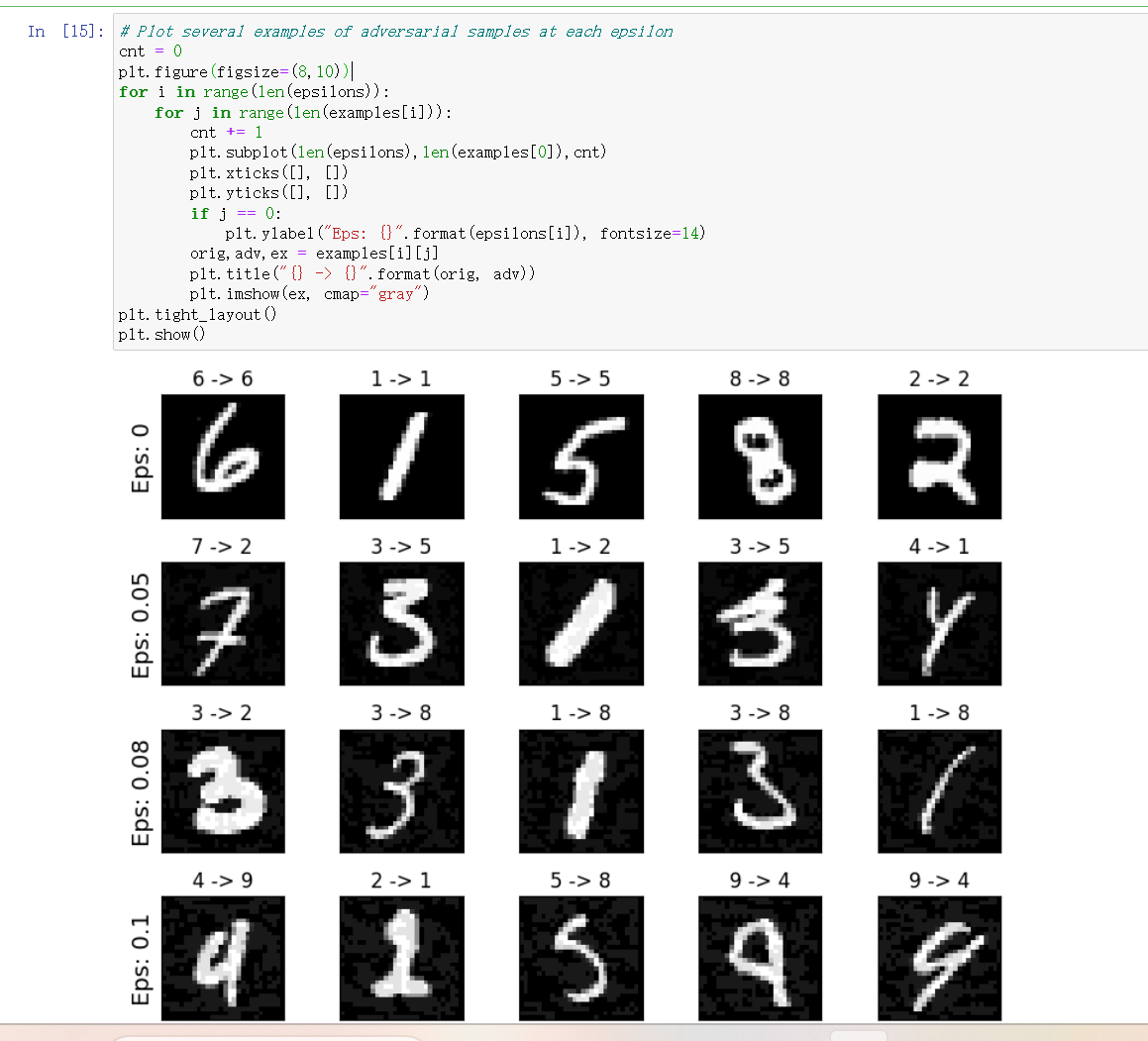
**结果分析**

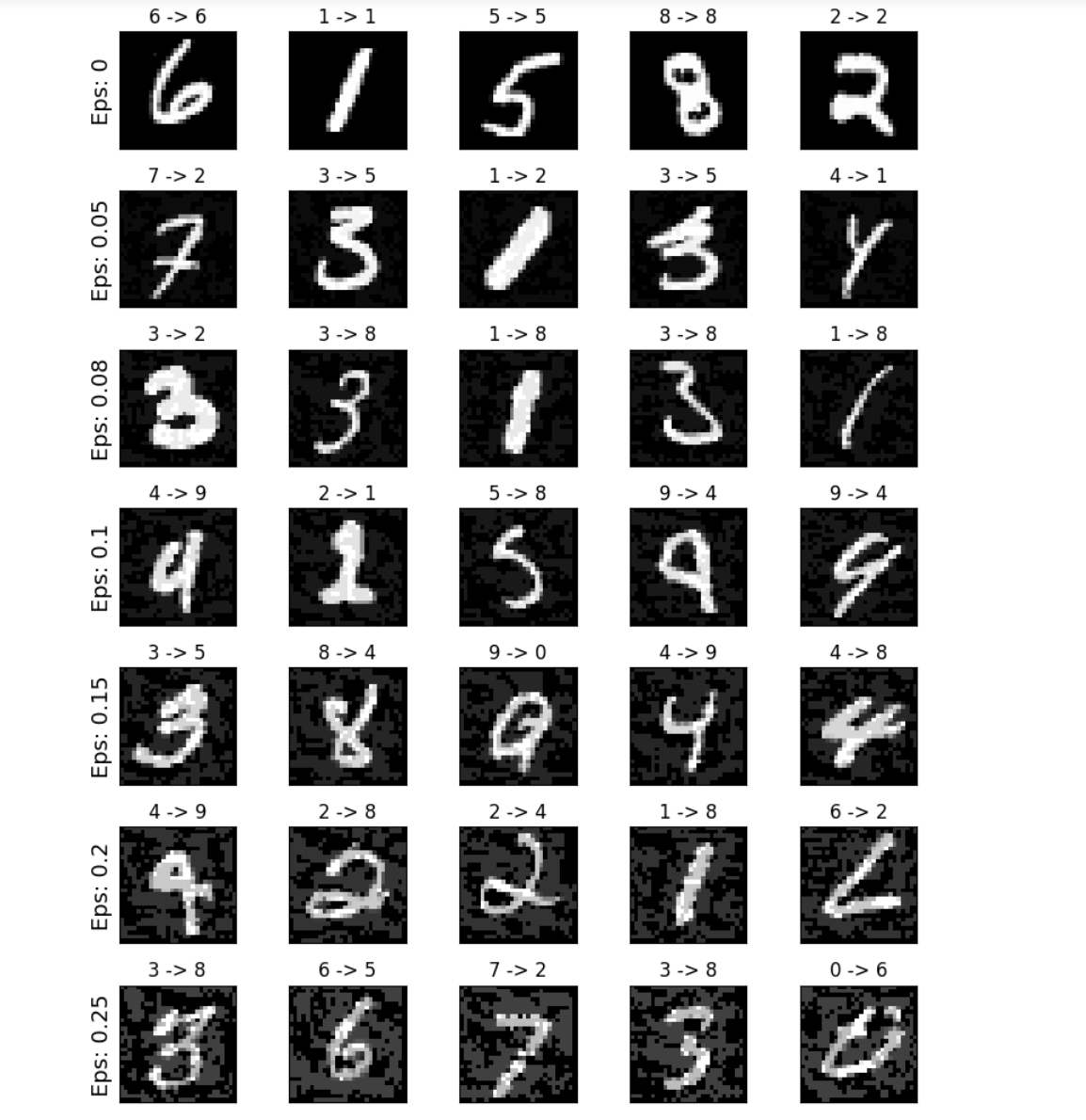
1. 准确性-扰动大小图



由图观察到，扰动越大，对图像的改变越大，模型识别的准确率越低，可以观察到，再没有进行对抗样本训练的时候，扰动大于0.1后模型准确率急剧下降

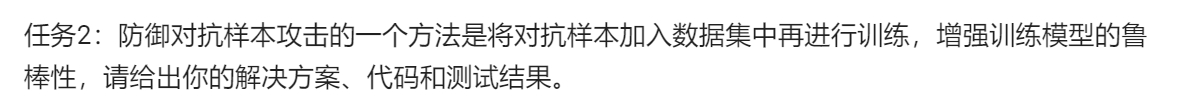
1. 对抗样本实例：





**任选一个扩展任务完成**

本人选择了任务2：



实现代码截图如下：

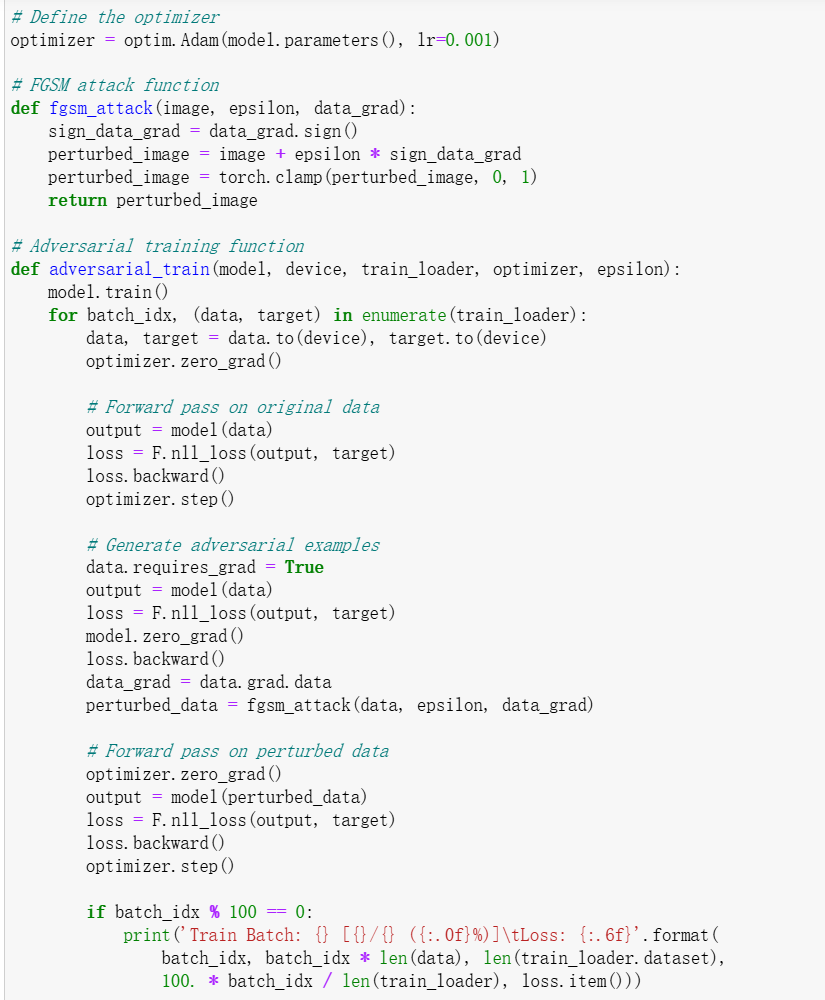
库的导入以及一些基本参数设置：

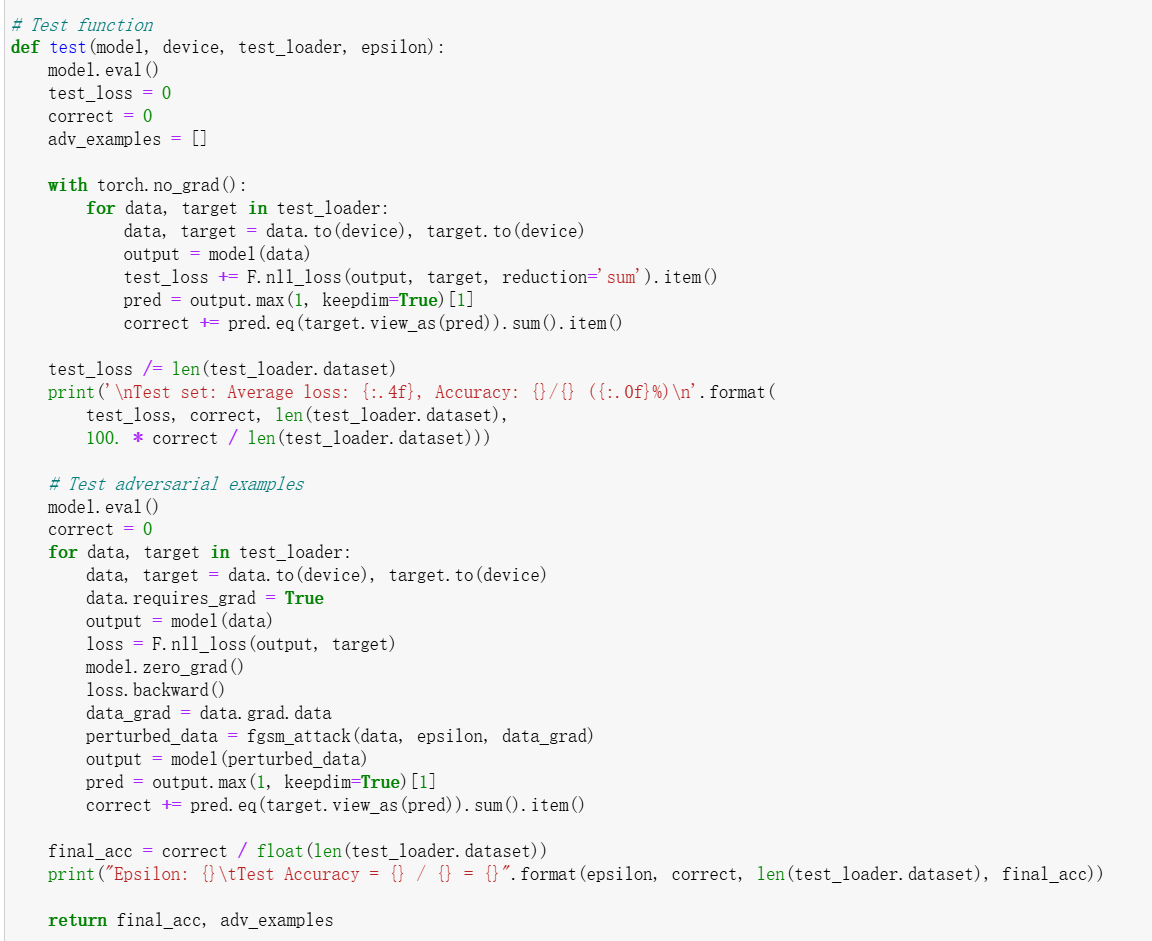


模型架构，模型权重，模型存储设备以及数据加载器设置：

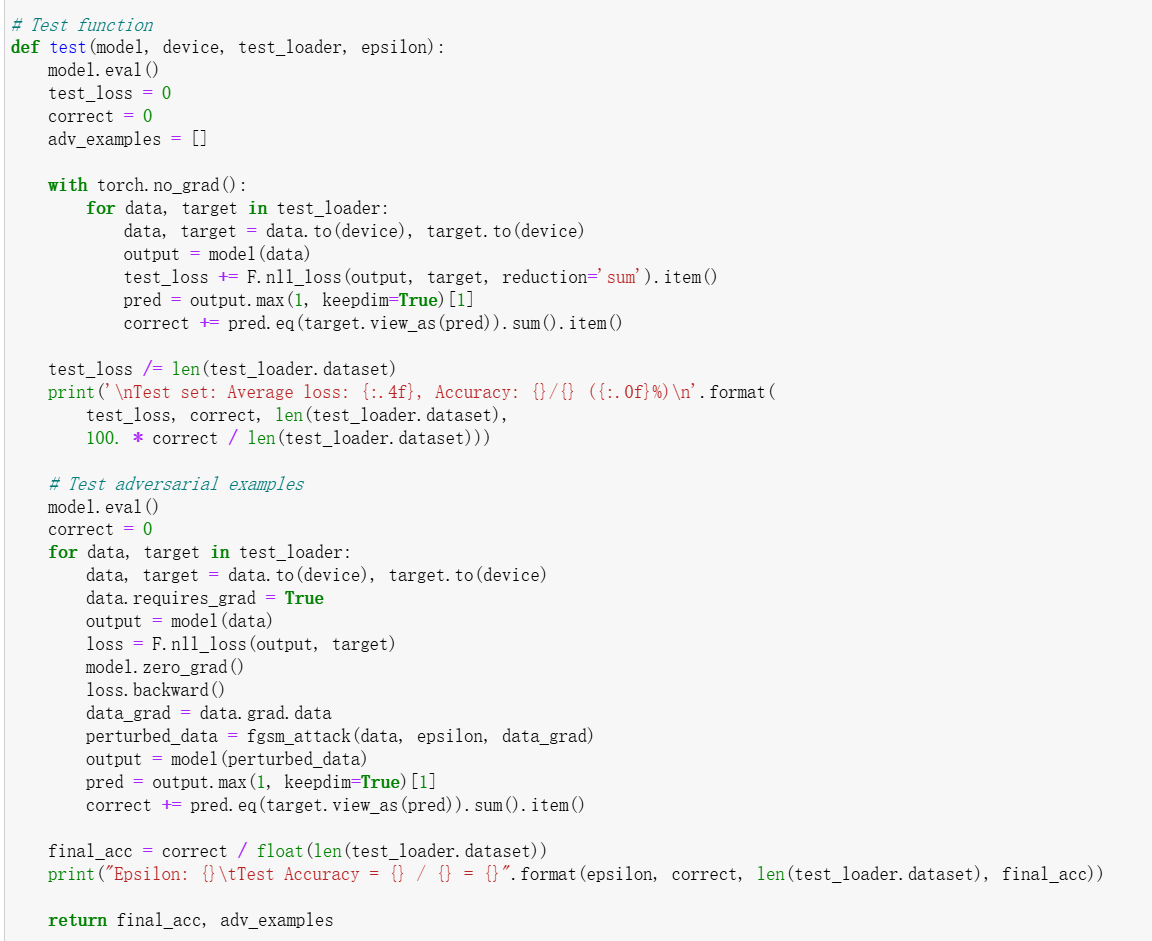


优化器，损失函数，训练函数（在训练中实时生成对抗样本参与训练），测试效果函数定义：

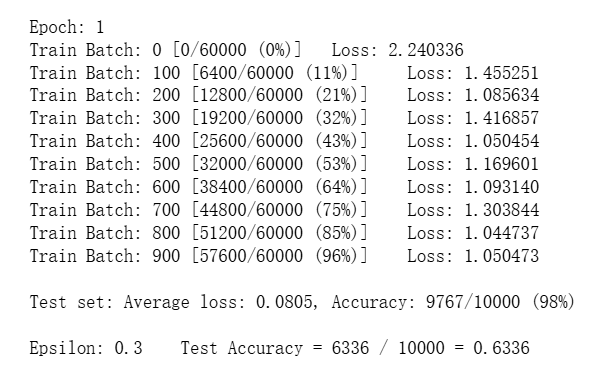


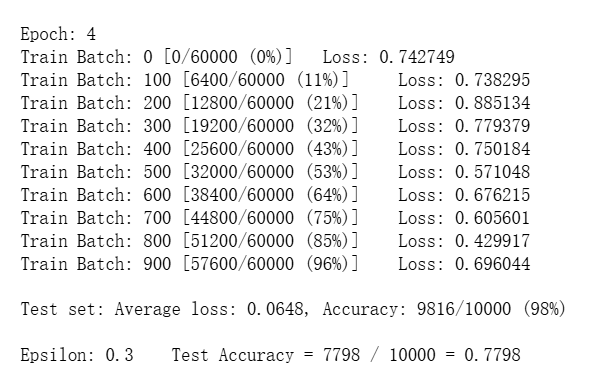


指定训练超参数，保存模型权重，可视化训练结果：

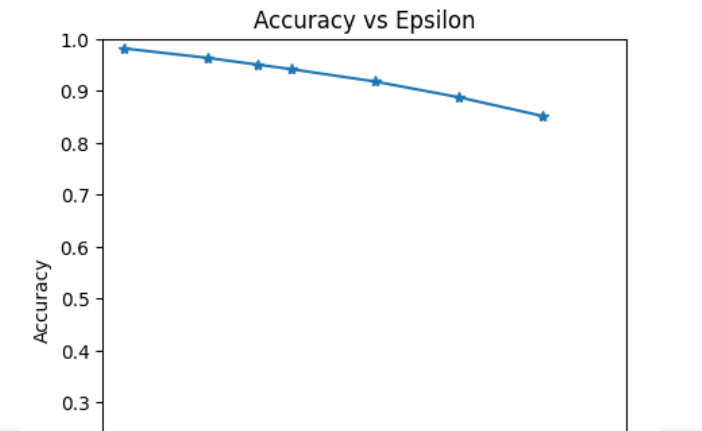


训练部分过程展示：





多个扰动训练后在测试上的可视化结果：



观察可知，加入对抗样本进行训练能有效提升模型对于对抗样本的识别能力，能有效提高模型对于对抗攻击的防御能力

# 二、遇到问题及解决方法

问题：对于对抗样本训练不知道如何产生对抗样本，何时用对抗样本作为数据参与训练

解决方法：查阅相关资料，咨询AI确立训练框架，两者结合解决对抗样本训练问题

# 三、对本次实验的建议

本次实验是唯一一个不用依赖于机房环境能够在本地环境很简单上手操作的实验，并且每一个实验也都只需要花课内时间便能完成，同时防火墙，SQL注入XSS蠕虫攻击等实验都很能增长编码以及相关知识的兴趣，是一门很好的课程