|  |
| --- |
| 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: | 梁鑫嵘 | | 学 号: | 200110619 | | 专 业: | 计算机科学与技 | | 日 期: | 2023年4月 |   **一、本次实验要求** |

1. 完成**4.3 FGSM 攻击函数**的代码补充，截图说明。



1. 分析**4.4 测试攻击效果函数** 的代码部分，说明每段代码的作用。

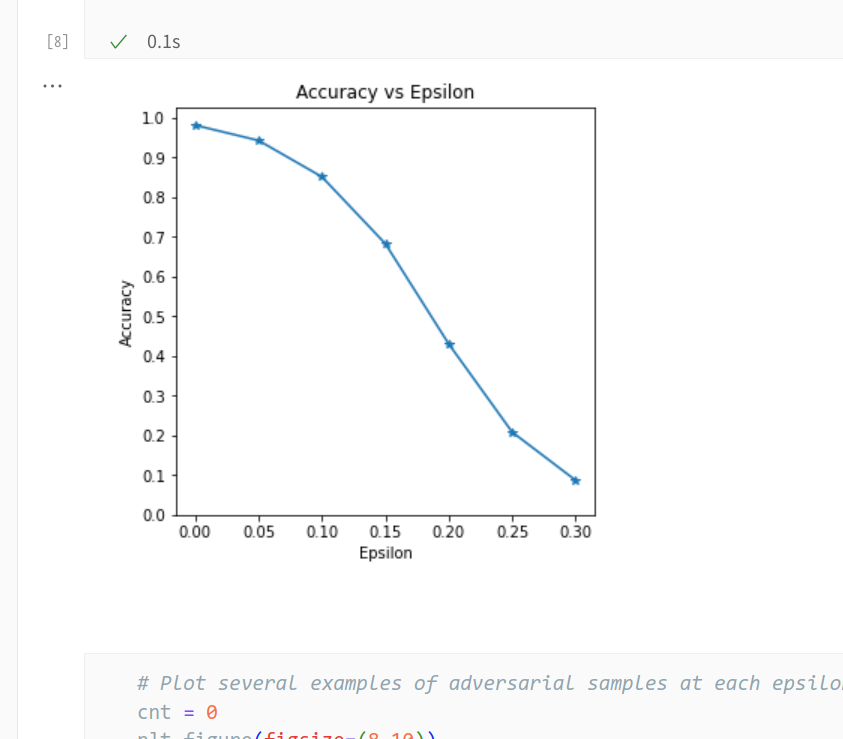
这段代码实现了一个用于对模型进行攻击的测试函数。

1. 函数名为test，接受以下参数：
   * model：要测试的模型
   * device：设备（如CPU或GPU）用于计算
   * test\_loader：测试数据集的数据加载器
   * epsilon：FGSM攻击中的扰动大小
2. 在函数内部初始化了计数器correct和存储对抗样本的列表adv\_examples。
3. 使用for循环遍历测试集中的所有样本。
4. 将数据和标签发送到指定设备。
5. 设置requires\_grad属性为True，以便在攻击中计算梯度。
6. 将数据通过模型进行前向传播，得到预测结果output和初始预测值init\_pred。
7. 如果初始预测与真实标签不一致，跳过此样本。
8. 计算损失loss，使用负对数似然损失函数（negative log-likelihood loss）。
9. 清零模型的所有梯度。
10. 在反向传播过程中计算模型的梯度。
11. 收集梯度数据data\_grad。
12. 调用fgsm\_attack函数进行FGSM攻击，生成扰动后的数据perturbed\_data。
13. 对扰动后的数据进行再分类，得到最终预测结果final\_pred。
14. 如果最终预测与真实标签一致，增加正确分类的计数器correct。如果epsilon为0且保存的对抗样本数量小于5个，则将对抗样本添加到adv\_examples列表中。
15. 如果最终预测与真实标签不一致，将对抗样本添加到adv\_examples列表中（数量不超过5个）。
16. 计算该epsilon下的最终准确率final\_acc。
17. 打印输出测试结果。
18. 返回最终准确率和对抗样本列表。

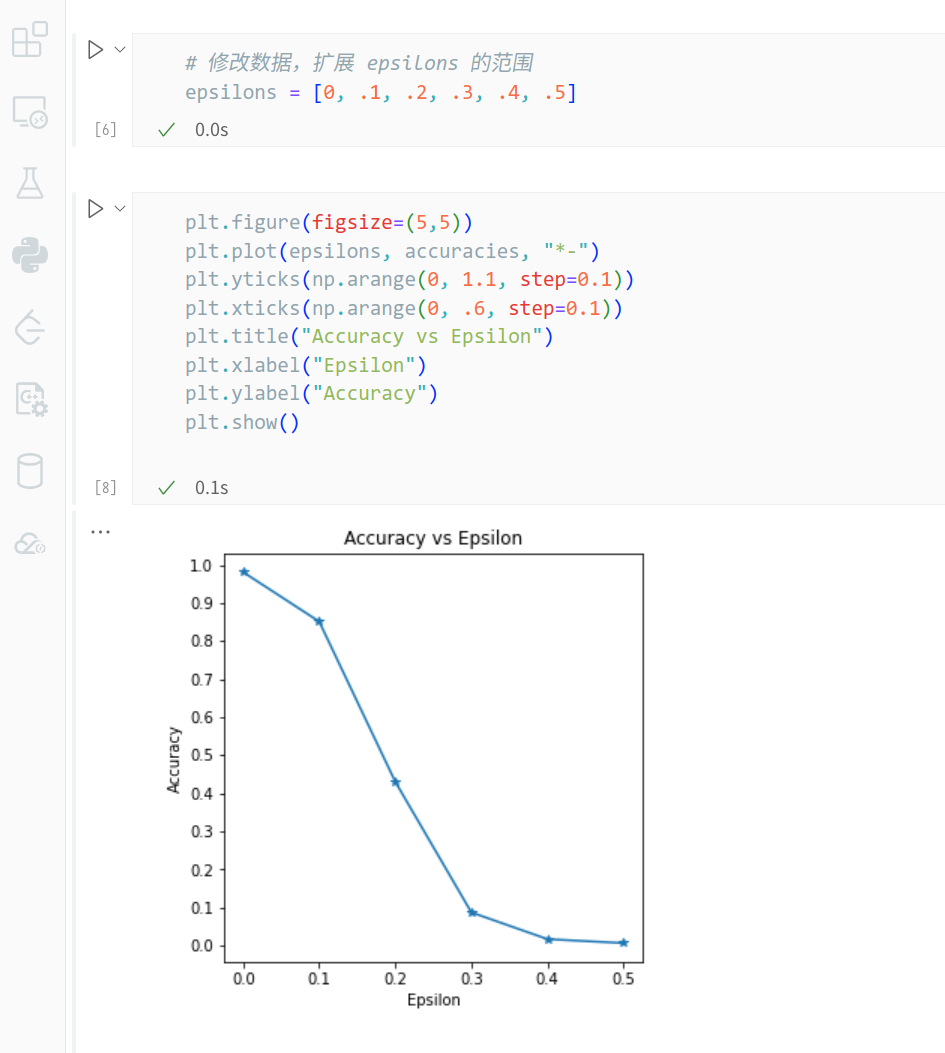
这段代码测试使用FGSM攻击后的模型在给定扰动大小epsilon下的准确率，并且保存一些对抗样本以供后续可视化分析。

1. 分别对 默认给出的epsilons = [0, .05, .1, .15, .2, .25, .3]和自行修改的epsilons 执行结果进行截图，并做简要说明。

对默认给出的epsilons执行的结果：



使用epsilons = [0, .1, .2, .3, .4, .5]的结果：



观察两个图像，得到以下结论：

1. 当espsilon增大，模型正确率减小
2. espsilon在0.1~0.3之间，Acc下降最快
3. espsilon在 >=0.3时，Acc下降不明显
4. 通过增大epsilon增强扰乱，能让正确率接近0，效果十分明显

二、网络与信息安全实验课程的收获和建议（**必填部分**）

*（关于本学期网络与系统实验的三个部分：系统安全，网络安全和AI安全，请给出您对于这三部分实验的收获与体会，给出评论以及****改进的建议****。）*

1. 系统安全部分

Meltdown Attack实验非常有趣，不过只能在实验室的老电脑上复现。

操作系统安全加固实验就只是对着PDF操作，对自身的提升比较有限。

1. 网络安全部分

从SeedLabs系列实验中了解了PKI公钥系统、TLS等技术，非常有用。

1. AI安全部分

了解了比较新颖和前沿的AI对抗样本攻击技术，对提升自己对AI安全方面的理解有很大帮助。