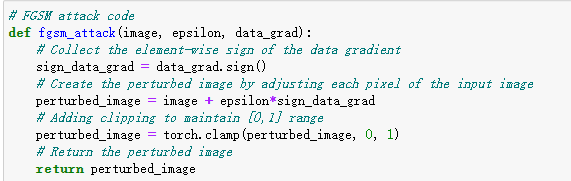
|  |
| --- |
| 哈尔滨工业大学(深圳) |
| **《网络与系统安全》 实验报告** |
|  |
| 实验七  对抗样本攻击 实验  学 院: 计算机科学与技术   |  |  | | --- | --- | | 姓 名: | 宗晴 | | 学 号: | 200110513 | | 专 业: | 计算机 | | 日 期: | 2023年4月 |   **一、本次实验要求** |

1. 完成**4.3 FGSM 攻击函数**的代码补充，截图说明。

FGSM攻击函数的代码如下图所示：

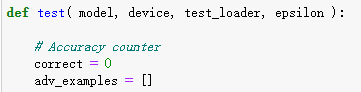


首先收集数据梯度的元素符号，通过sign()函数实现。然后通过调整输入图像的每个像素来创建扰动图像，具体实现是对每一个像素加上epsilon\*sign\_data\_grad的扰动。然后添加剪切以维持[0,1]范围，通过torch.clamp函数实现。最后返回被扰动的图像。

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。

测试攻击效果函数的代码部分如下图所示。

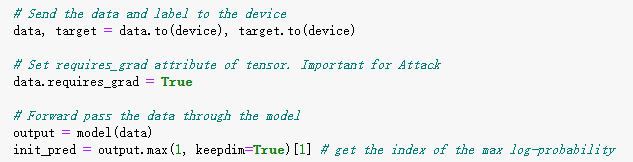
首先，设置精度计算器：



然后循环遍历测试集中的所有示例：

C:\Users\lenovo\Documents\WeChat Files\wxid_xgzr63lnxlp912\FileStorage\Temp\1686214059574.png

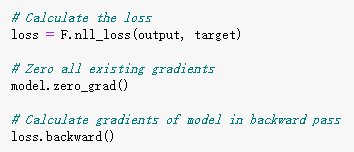
把数据和标签发送到设备，设置张量的requires\_grad属性，并通过模型前向传递数据：



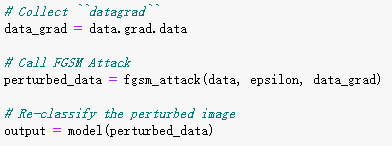
如果初始预测是错误的，不打断攻击，继续：

C:\Users\lenovo\Documents\WeChat Files\wxid_xgzr63lnxlp912\FileStorage\Temp\1686214132740.png

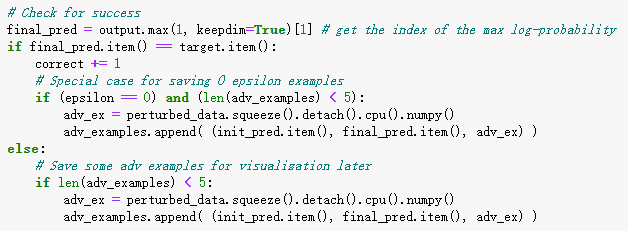
计算损失，将所有现有的梯度归零，将损失反向回传：



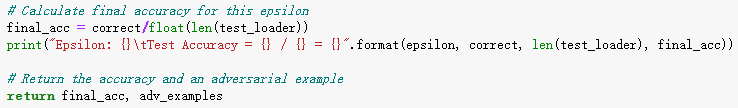
收集datagrad，调用上述的fgsm\_attack函数进行攻击，然后重新分类受到扰动的图像：



检查是否成功。同时保存0 epsilon示例的特例，以及一些其它epsilon的示例用于后续的可视化：

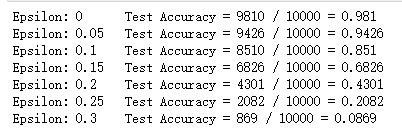


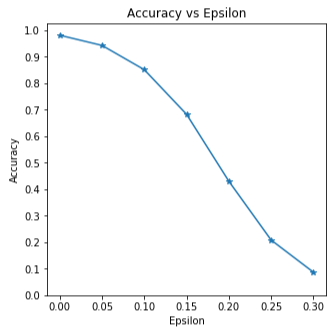
循环结束，计算当前epsilon下的最终准确率，返回准确率和对抗性示例：



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。

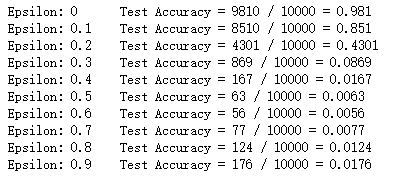
epsilons = [0, .05, .1, .15, .2, .25, .3] 的结果如下图所示：

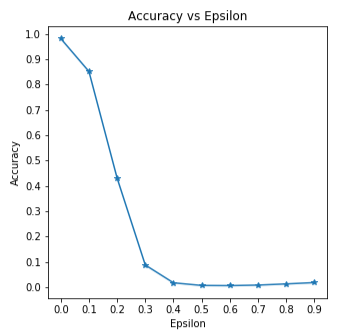


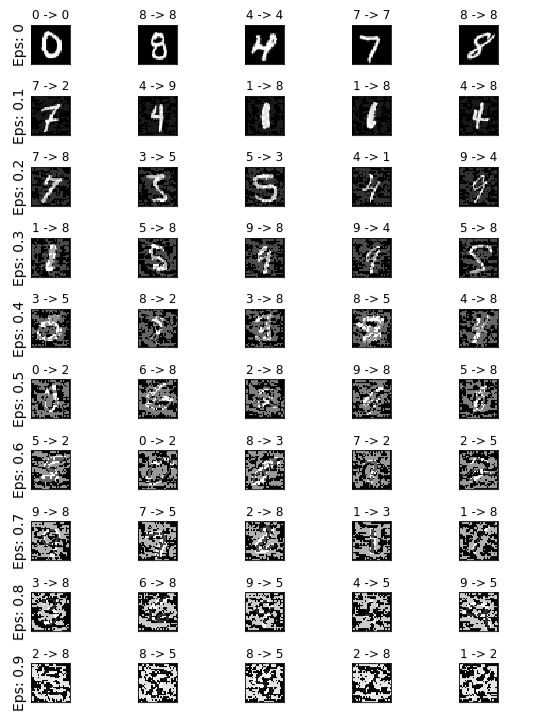




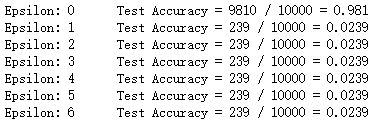
epsilons = [0, .1, .2, .3, .4, .5, .6] 的结果如下图所示：

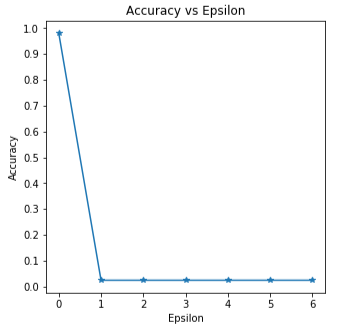


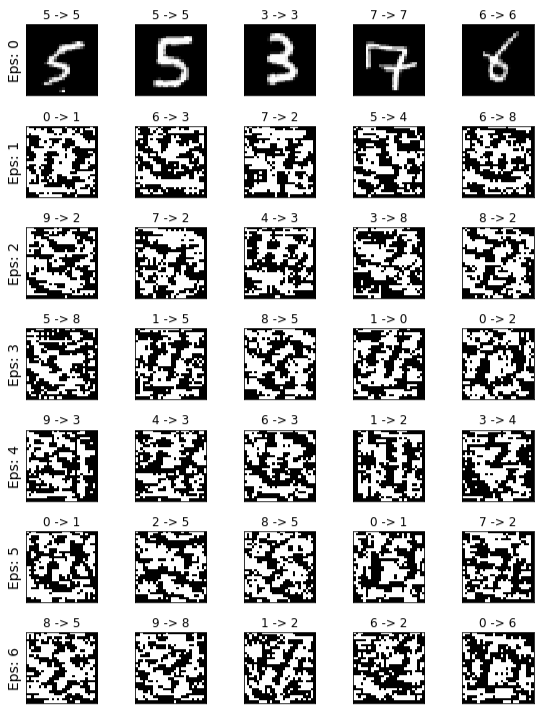




epsilons = [0, 1, 2, 3, 4, 5, 6] 的结果如下图所示：







简要说明：

由上述实验可以看出，随着epsilon从0开始逐渐增大，模型的准确率在逐渐降低；当epsilon在0到0.3之间时，准确率发生骤降；当epsilon = 0.3时，准确率已经小于0.1；当epsilon在0.3到0.6之间时，准确率依然在小幅度下降；当epsilon = 0.6时，准确率最低，为0.0056；然后当epsilon继续不断增大到1时，准确率会有一段小幅度的提升；最终当epsilon大于1时，准确率维持在0.0239保持不变。

并且从给定的示例中可以看出，当epsilon在0到0.2之间时，经过扰动后的图像还十分清晰，人眼可以准确地判断出来；当epsilon在0.2到0.7之间时，经过扰动后的图像变得较为模糊，但通过人眼依然能依稀辨认出来；当epsilon大于0.7时，图像受扰动之后变化较大，通过人眼已经无法再辨认出来。

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****。）*

本学期的实验中系统安全和AI安全相对来说较为简单，网络安全部分相对更为困难，我在上面花费了较多的时间。

系统安全部分的指导书较为详细，按照所述步骤进行实验能够比较顺利地完成。网络安全部分，由于第一次接触相关的概念，所以实验的时候会产生一些疑问，有的时候可能指导书上提到了，但是因为对相关概念的印象并不深，所以可能没有注意到指导书的提示，比如要在容器A中进行操作，然后再到容器B中进行操作，第一次进行相关实验的时候没有注意到要换容器，所以也出现了很多问题。AI安全部分的实验相对也比较简单，因为实验提供了代码框架，需要做的只是补充完整一些函数，同进行调参并分析结果，完成的比较顺利。总之，在完成网络与信息安全的所有实验之后，我对于网安这门课程有了更深入的认识，网络与信息安全覆盖了各个领域。

改进的建议就是网络安全部分的指导书可以再详细一些，比如标注出一些重点，或者是给出一些错误示例，让同学们知道发生这样的错误可能是什么原因产生的，从而可以自己进行纠正，而无需每个同学遇到错误都向老师询问，也可以适当减轻老师们的负担。以及最后一部分的AI安全实验可以略微加大难度，目前来说有些过于简单了。