**智能应用建模课程报告**

**丁豪，男，南京大学人工智能学院，181220010**

1. 对抗样本的背景
2. **什么是对抗样本**

对抗样本是一种出于破坏神经网络正确识别能力的目的，而在原始输入数据上加上可以容忍的特定小噪声，以此来极大影响神经网络最终的输出结果的方法。其针对目标可以分为白盒神经网络与黑盒神经网络，大部分对抗方法具有可复现性与一定的迁移性，并且由于神经网络本身的复杂结构，防御对抗样本的难度相对较大。

1. **为什么我们要学习对抗样本**

正所谓“知己知彼百战不殆”，充分理解我们频繁使用的神经网络很容易受到对抗样本攻击的事实，更有利于我们安全合理地使用神经网络进行人工智能开发。针对不同的对抗方法，研究相应的防御措施，也是人工智能安全性提升的重要手段。

1. 攻击方法

在本次作业中使用了 i-FGSM，白盒C&W，黑盒ZOO\_Newton 三种方法，将在下一节实现与评估中详细讲解。

1. 实现与评估

注：所有代码均已包含在压缩文件内

1. **训练模型**

使用keras进行mnist数据集模型训练，使用了两个卷积、池化组和两个全连接层，外加中间的一些relu激活，一个拉直，最终的softmax。Batch\_size设定为128，训练轮数为5。所有数据均使用/255来实现[0,1]。

最终在纯净mnist数据集上的识别正确率达到了：98.99%

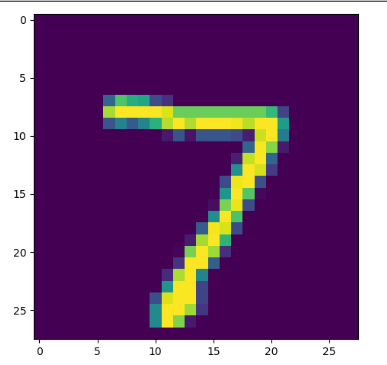
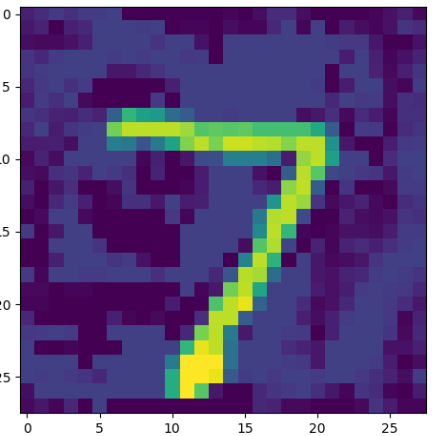
1. **i-FGSM**

本方法采用“迭代式快速梯度方向下降”方法进行优化，直接求解目标方程与原始输入数据的梯度，并对输入数据每一个元素根据梯度正负方向下降设定好的尺寸，不断迭代直到满足最终目标--识别为错误的类或目标类。在非目标攻击时，目标方程即所有“不正确”分类中概率最大的类的概率值。在目标攻击中，目标方程就是“目标”类的概率值。为防止与原图差别过大，设定最大偏离量s并使用[x-s,x+s]来约束每次更新后的数据。由于梯度下降可能会产生超过[0,1]的非法数据，因此又使用[0,1]boxing来规约每一次迭代后的最终数据。

为了节省时间，限定每张图的最大迭代次数为30，偏移量最大为0.2。得到的无目标攻击成功率：99.4%，目标攻击成功率：88.7%。平均l2 norm为：2.235



攻击的示例图片如下：

**（原始图，识别为7） （攻击后的图，识别为5，l2 norm为1.97）**

1. **白盒C&W**

C&W方法的几个公式如下，中心思想是同时优化与原图的差距和攻击产生的差异，通过合理调试c的取值就可以达到这种效果。







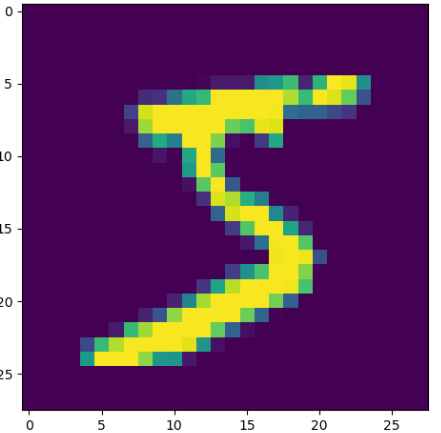
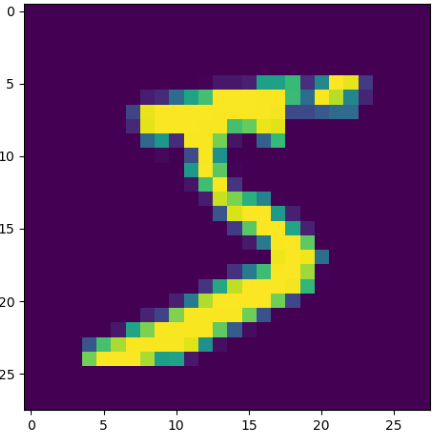


通过限定p为2，实现针对L2的攻击,使用Adam来进行优化。公式中的-k选择为0，c从0.001开始使用对数尺度上的线性搜索来确定最终值。

由于没有指定攻击图片与原图的差距，所以攻击总是可以成功，成功率为100%。平均l2 norm为（图片已经处理成[0,1]）



攻击示例图片如下：

（原图，识别为5） （L2攻击后，识别为3，l2 norm 1.42）

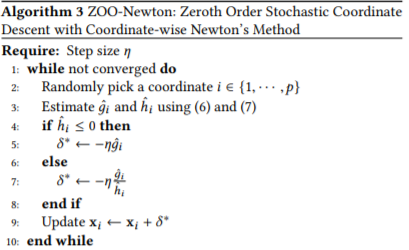
1. **黑盒C&W ZOO\_Newton方法**







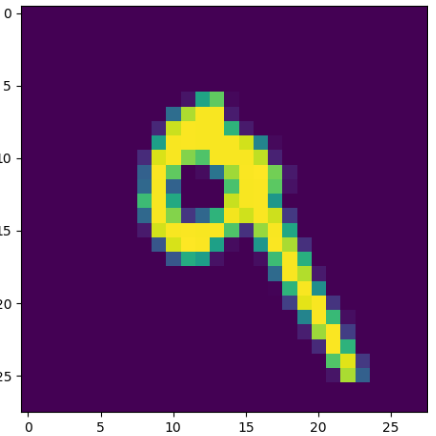
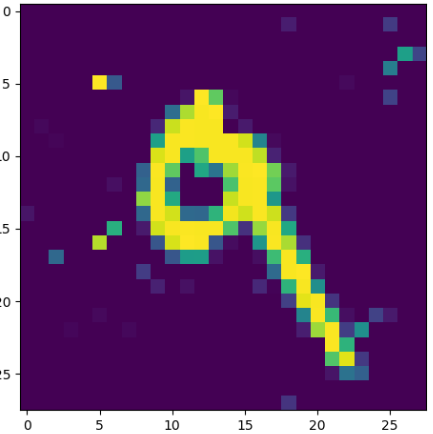




使用python完全实现上述算法即可。手动调整c的大小来兼顾l2 norm与攻击效果，最终确定为0.003。此算法的一大亮点在于完全不需要知道网络结构，直接通过定义与简单近似来求得所需minimize函数的一阶与二阶微分，理论上可以运用于任何类型的黑盒网络攻击。

对抗样本的平均l2 norm为：(已经处理成[0,1])



（原图，识别为9） （黑盒CW对抗后，识别为4, l2 norm 3.96）