上机实验二：对抗样本攻击与防御

任务要求：在实验一 MNIST 数据集 10 分类手写体数字识别的基础上，使用 FGSM (Fast Gradient Sign Method) 方法对卷积神经网络进行**（i）非定向（untargeted）对抗样本攻击**、**（ii）定向（targeted）对抗样本攻击**、以及**（iii）防御**。提交代码以及实验报告，实验报告中附上定向和非定向情况下的对抗样本图片（每一类别给出一张示例图片即可）、定向和非定向攻击成功率以及防御效果，并进行适当分析。

具体要求：

1. 非定向攻击要求通过生成对抗样本使得神经网络产生任意不同于原图片真值 i 的输出，即对抗样本使得神经网络输出不等于 i 即可；
2. 定向攻击要求生成对抗样本使得真值为 i 的图片被神经网络误判为 i+1（0 朝 1 攻击，1 朝 2 攻击，以此类推，9 朝 0 攻击）.
3. 探索不同阈值$\varepsilon = \{1,10,20,50\}/255$下的攻击效果.
4. 采用对抗训练方式进行防御，即将非定向攻击生成的对抗样本进行准确的真值标记，联合原有训练数据集在现有神经网络的基础上进行对抗重训练。
5. 通过对经过对抗训练的神经网络再次进行非定向对抗攻击，计算该次攻击成功率来测试防御效果。
6. 做好代码注释。

提交：

1. 将代码和实验报告打成压缩包，以“姓名+学号”命名文件，比如“张三+2019XXX”，发送到邮箱：xzli@hust.edu.cn
2. 截止时间为 6 月 12 号下午 5:00。