图像分类模型的对抗攻击和对抗训练 作业结果总结

金奕同 学号：2101990330

1. 白盒攻击：

* 分类器在test集上的准确率： **91.21%**

Fashion MNIST图像分类器源码请见：‘./白盒攻击/whitebox\_code/’ 目录。

* 白盒攻击的成功率：**59.10%**

白盒攻击源码请见: ‘./白盒攻击/whitebox\_code/whitebox\_attack.py’。

* 白盒攻击所得的对抗样本存于 ‘./白盒攻击/whitebox\_attack\_data/adversarial\_samples\_1k.pkl’文件中，用于之后的对抗训练。10组原图像与攻击成功的图像对比，以及分类器对他们的类别判断请见：‘./白盒攻击/whitebox\_adversarial\_samples/’目录中。

1. 黑盒攻击：

* 黑盒攻击的成功率：**17.90%**

黑盒攻击源码请见: ‘./黑盒攻击/ code/blackbox\_attack.py’

* 10组原图像与攻击成功的图像对比，以及分类器对他们的类别判断请见：‘./黑盒攻击/blackbox\_adversarial\_samples/’目录中。

1. 对抗训练：

* 对抗训练得到的新分类器在test 集上的准确率：**91.78%**

VS 对抗训练得到的旧分类器在test 集上的准确率：**91.21%**

将对抗样本掺入训练集后训练新模型的源码请见：‘./对抗训练/advTrain\_code/adversarial\_train.py’。运行‘./对抗训练/advTrain\_code/test.py’文件，即可复现新分类器在test 集上的准确率，并将新模型在test集上预测正确的1000个随机样本存入‘./对抗训练/advTrain\_attack\_data/adv\_model\_correct\_1k.pkl’文件中，以用于接下来的白盒、黑盒攻击。

* 在使用相同的迭代上限的条件下：

白盒攻击在新分类器上的攻击成功率：**48.10%**

VS 白盒攻击在旧分类器上的攻击成功率：**59.10%**

白盒攻击源码请见：‘./对抗训练/advTrain\_code/whitebox\_attack.py’

* 在新、旧分类器上各取1000个预测正确的样本，用额外的模型“extraCNN”上进行白盒攻击得到攻击样本，使用样本迁移的方法对新、旧分类器进行黑盒攻击，在使用相同的迭代上限的条件下：

黑盒攻击在新分类器上的攻击成功率：**29.30%**

VS 黑盒攻击在旧分类器上的攻击成功率：**31.30%**

黑盒攻击源码请见：‘./对抗训练/advTrain\_code/blackbox\_attack.py’

额外模型及其训练请见：‘./对抗训练/advTrain\_code/extra\_model.py’和‘./对抗训练/advTrain\_code/train\_extra\_model.py’文件。

* 在新分类器上，白盒攻击成功的10组样本，请见：

‘./对抗训练/advTrain\_whitekbox\_samples/’ 目录。

* 在新分类器上，黑盒攻击成功的10组样本，请见：

‘./对抗训练/advTrain\_blackbox\_samples/’ 目录。