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。

2. 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。

3. 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。