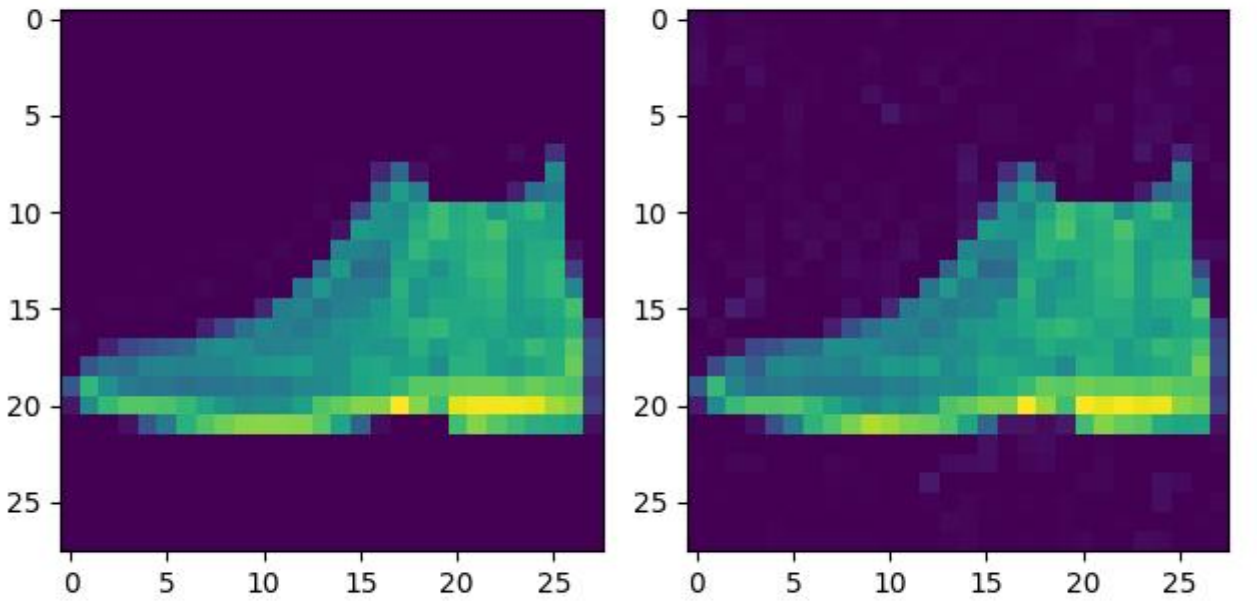
# 图像分类模型的对抗攻击和对抗训练

**刘家铭 2301210643**

所有代码在mycode文件夹中，白盒攻击相关图片在whitebox文件夹中，黑盒攻击相关图片在blackbox文件夹中。

## 白盒攻击

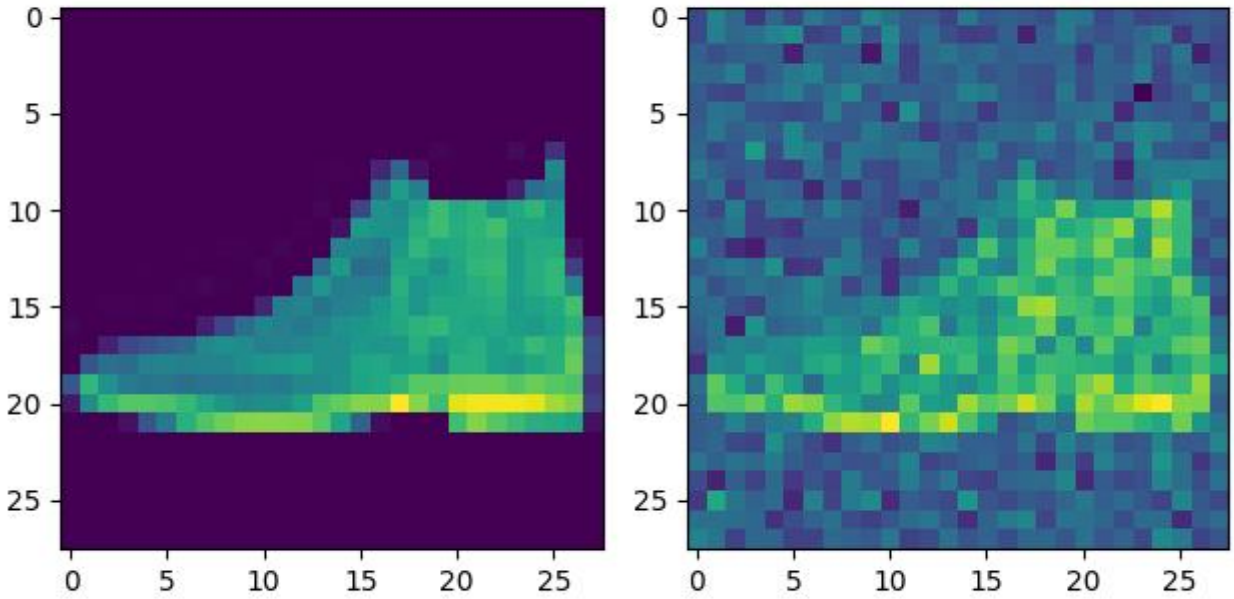
训练了一个Fashion MNIST上的图像分类模型，在mymodel文件夹中有mymodel的checkpoint。模型定义在model.py文件中。白盒攻击的代码在wbox.py文件中。




左图为原始样本，右图为攻击后的样本。原始标签为AnkleBoot，攻击后标签为T-shirt。

## 黑盒攻击

对给定模型进行了黑盒攻击，具体的攻击方法相关代码在bbox.py文件中。




左图为原始样本，右图为攻击后的样本。原始标签为AnkleBoot，攻击后标签为T-shirt。

## 对抗训练

加入对抗样本之后训练了一个新的模型，在mymodel文件夹中有newmodel的checkpoint。分别对旧模型和新模型进行白盒攻击和黑盒攻击，对自己模型进行黑盒攻击的代码在bbox_m.py中。

## 实验结果

自己训练的分类器在test集上的准确率为91.05%。

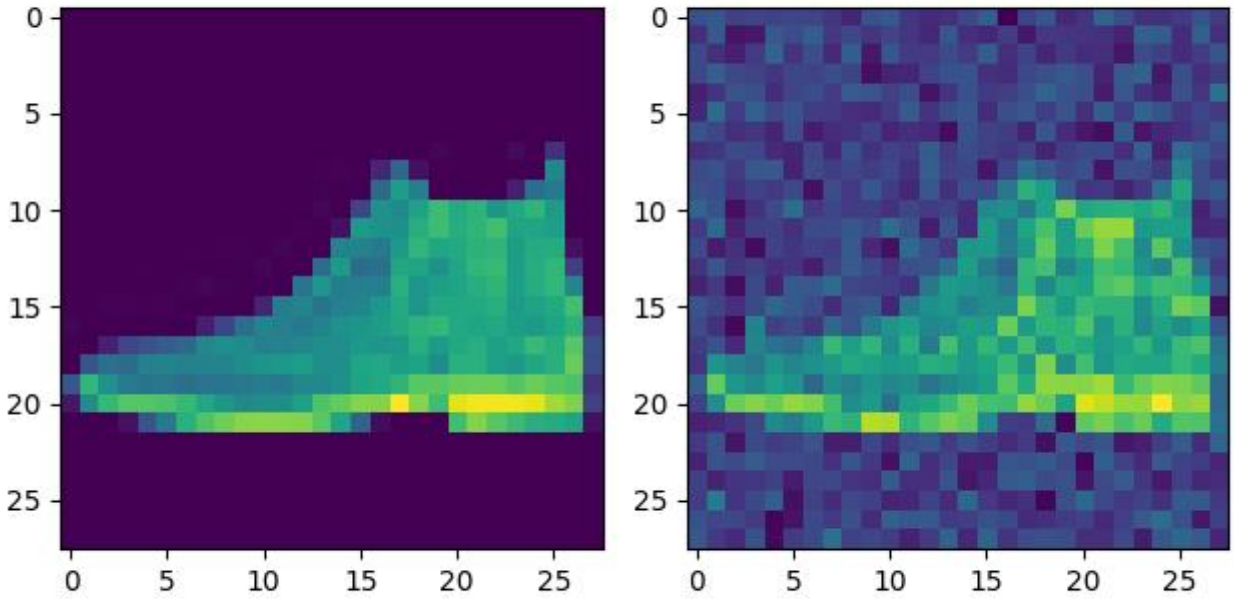
对待攻击模型的黑盒攻击成功率为54.80%。

对抗训练后的新分类器在test集上的准确率为89.38%。

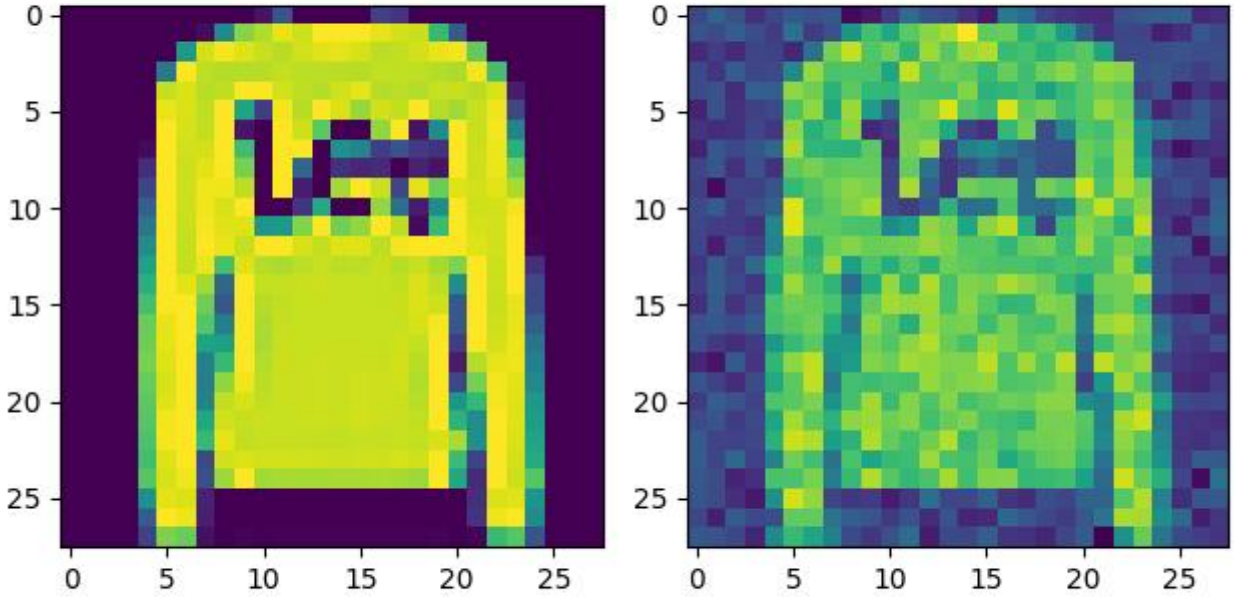
对旧模型的白盒攻击成功率为96.10%。

对新模型的白盒攻击成功率为90.10%。

对旧模型的黑盒攻击成功率为54.30%，一个成功的样本如下，左图为原始样本，右图为攻击后的样本。原始标签为AnkleBoot，攻击后标签为T-shirt。




对新模型的黑盒攻击成功率为25.20%，一个成功的样本如下，左图为原始样本，右图为攻击后的样本。原始标签为Pullover，攻击后标签为Dress。


0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25