DL第二次作业-图像分类模型的对抗攻击和对抗训练

2019-04-08

1801210116-张琦

step 1 训练一个 Fashion MNIST 上的图像分类模型

- 数据集： Fashion MNIST
- 网络模型：ResNet18
- 超参数：epoch = 20；learning_rate=0.001（每5个epoch减小一半），batch_size=100
- 测试集上正确率：93%
- 开发工具：python 2.7+pytorch + gpu
- 相关源码：network.py(网络结构）；train.py(模型训练+测试)

step 2 对选出的图像进行白盒攻击

- 攻击算法：梯度下降法
- 攻击成功率：70%
- 相关源码：whiteboxattack.py
- 结果展示：

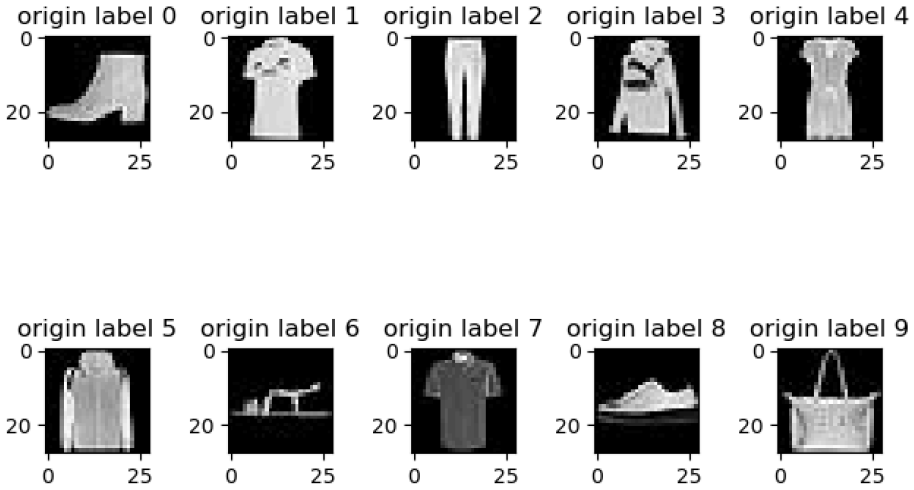

图1 白盒攻击原始图片

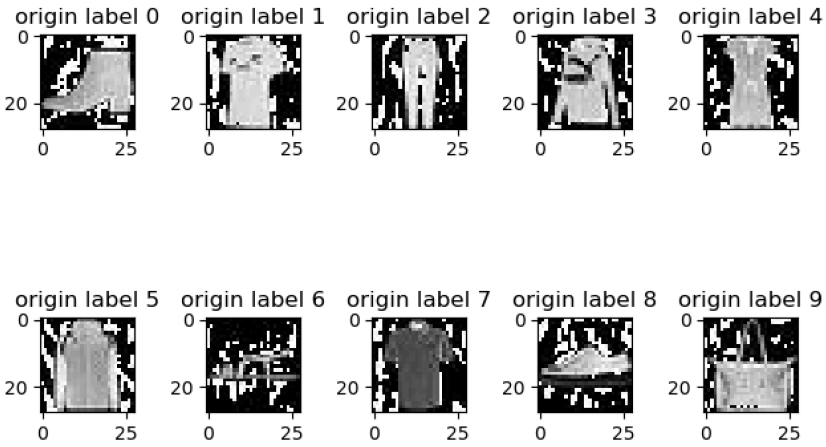

图2 白盒攻击攻击成功图片

step 3 对选出的图像进行黑盒攻击

- 攻击算法：MCMC 采样
- 攻击成功率：25%
- 相关源码：blackboxattack.py
- 结果展示：

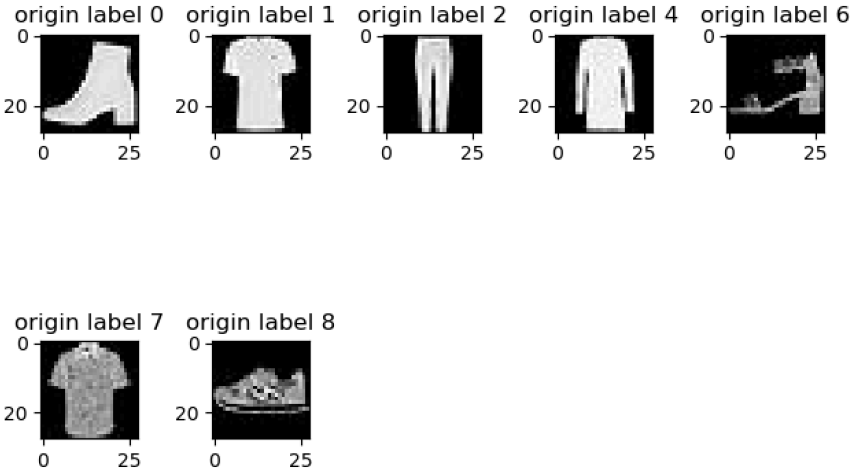

图3 黑盒攻击原始图片

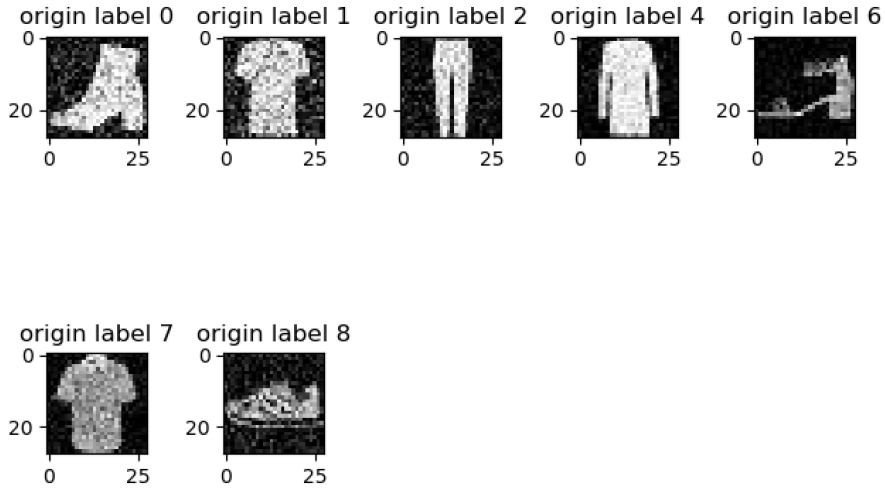

图4 黑盒攻击成功攻击的图片

step 4 将对抗样本掺入训练集中，重新独立训练一个分类器

- 相关源码：train_adv.py
- 测试集上正确率：92%

step 5 在新分类器上重复白盒攻击

- 攻击成功率：56%
- 相关源码：WA_Newmodel.py

step 6 在新分类器上重复黑盒攻击

- 攻击成功率：3%
- 相关源码：BA_Newmodel.py

附：训练结果截图

```
Epoch [20/20], Step [300/600] Loss: 0.1318
Epoch [20/20], Step [400/600] Loss: 0.1299
Epoch [20/20], Step [500/600] Loss: 0.1249
Epoch [20/20], Step [600/600] Loss: 0.1488
Accuracy of the model on the test images: 93 %

Process finished with exit code 0
```

图5 原始分类器训练结果

```
/home/xxxfrank/anaconda2/bin/python /home/xxxfrank/zqHomework/DLhw/whiteboxattack.py
model is loaded
700  pictures was attackted attacking rate :70 %

Process finished with exit code 0
```

图6白盒攻击结果

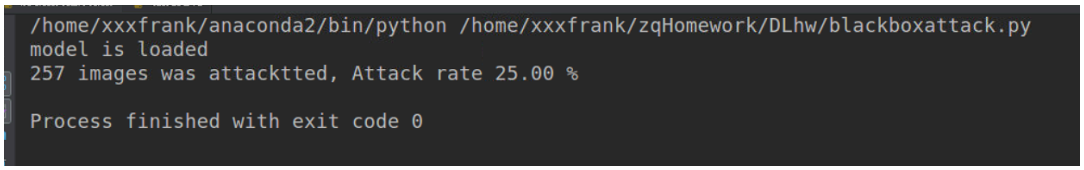

图7 黑盒攻击结果

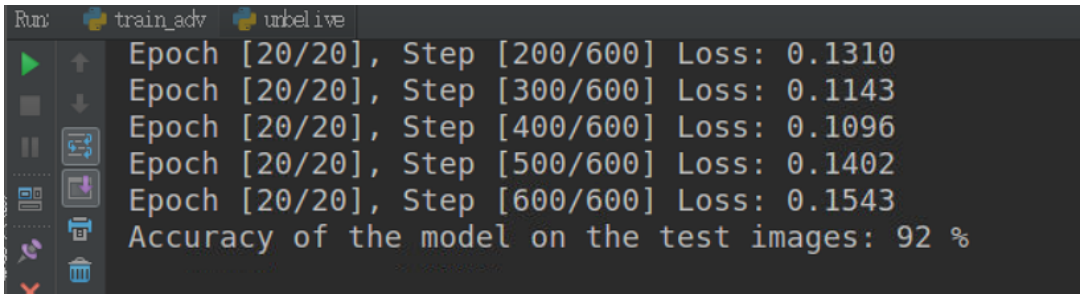

图8 加入对抗样本后训练的网络结果

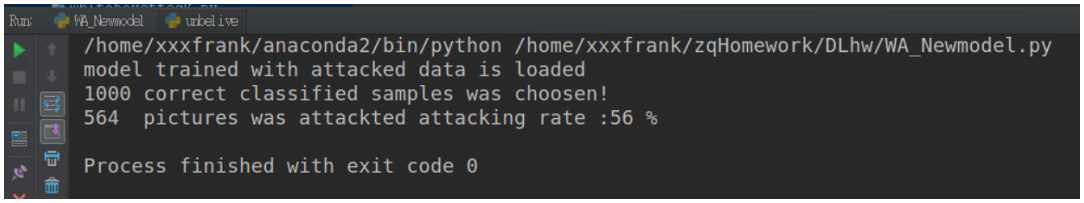

图9 新网络模型上百盒攻击成功率，较原来下降14个点

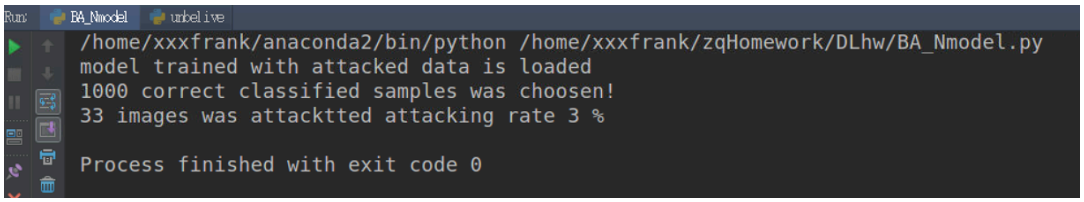

图10 新网络上黑盒攻击成功率，较原来下降22个点（因为对抗样本采用的黑盒攻击产生的对抗样本）