对抗训练报告

信息科学技术学院
智能科学系
彭晗 1801213805

2019 年 3 月 27 日

1 新分类器训练过程

在原始模型的训练集上进行白盒攻击，得到 Success rate=1652/55130 (3.0%), 其中 55130 为训练数据集中分类正确的样本，故可以用于对抗生成。

将 1652 个对抗样本加入到训练数据集中，得到含有 61652 个样本的新训练集。使用这些样本对原始模型进行训练，得到 Average loss=0.2722, Accuracy=9082/10000(91%)

2 准确率

2.1 新分类器

测试数据集上得到 Average loss=0.2722,Accuracy=9082/10000(91%)

2.2 旧分类器

测试数据集上得到 Average loss=0.2844,Accuracy=8989/10000(90%)。

相比旧分类器，新分类器性能略有提升。

3 白盒攻击

对抗算法采用较为基础的 FGSA(Fast Gradient Sign Attack) 算法，该算法设计于论文 Explaining and Harnessing Adversarial Examples 中。设置 epsilon 为 0.3，在测试数据集中进行实验。

3.1 新分类器

在 9082 个可供对抗生成的分类正确的样本中，共有 410 个样本对抗成功，成功率为 4.5%，选取十组样本进行展示。

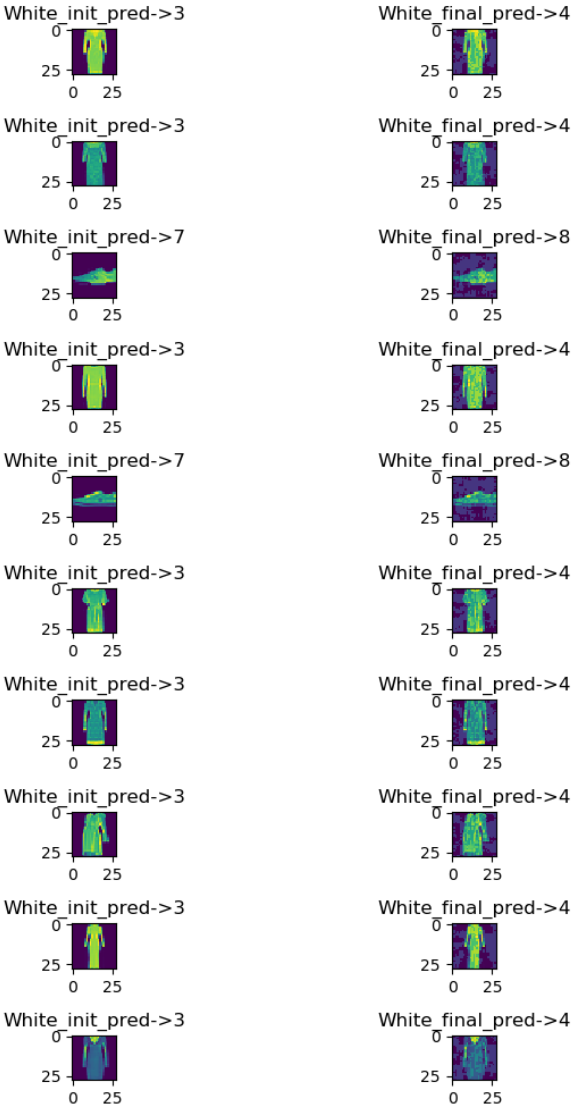

Figure 1: 10 组对抗样本（白盒攻击）

3.2 旧分类器

在 8989 个可供对抗生成的分类正确的样本中，共有 277 个样本对抗成功，成功率为 3.1%。由于数据较小，攻击成功率出现相反的结果应属正常。

4 黑盒攻击

黑盒攻击方式和任务二一致。

4.1 新分类器

在 9082 个可供对抗生成的分类正确的样本中，共有 420 个样本对抗成功，成功率为 4.6%，选取十组样本进行展示。

4.2 旧分类器

在 8989 个可供对抗生成的分类正确的样本中，共有 138 个样本对抗成功，成功率为 1.5%。由于数据较小，攻击成功率出现相反的结果应属正常。

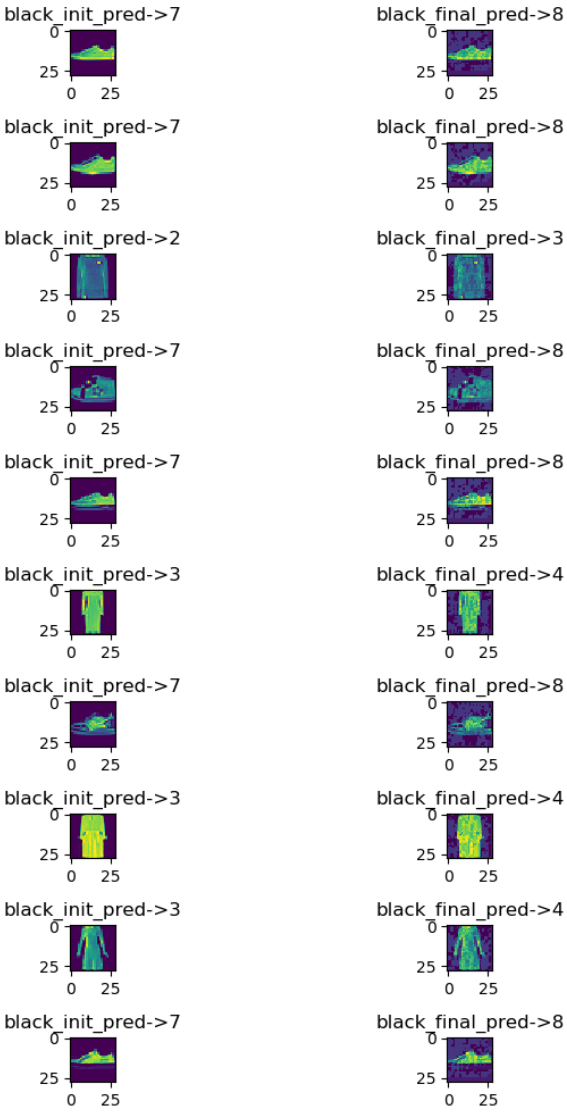

Figure 2: 10 组对抗样本（黑盒攻击）