白盒攻击报告

信息科学技术学院
智能科学系
彭晗 1801213805

2019 年 3 月 27 日

1 模型

分类器由两层卷积层、两层全连接层以及两层池化层组成，第一层卷积层使用 20 个 5*5*1 的卷积核，第二层使用 2*2 核进行 max 池化，第三层卷积层使用 50 个 5*5*20 的卷积核，第四层使用 2*2 核进行 max 池化，第五层使用全连接层至 500 维并使用 relu 作为激活函数，第六层使用全连接层至 10 维并使用 softmax 作为激活函数。

在 10000 个样本的测试集中，有 8989 个样本预测正确，准确率约为 90%。

2 对抗攻击

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。

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%，选取十组样本进行展示。

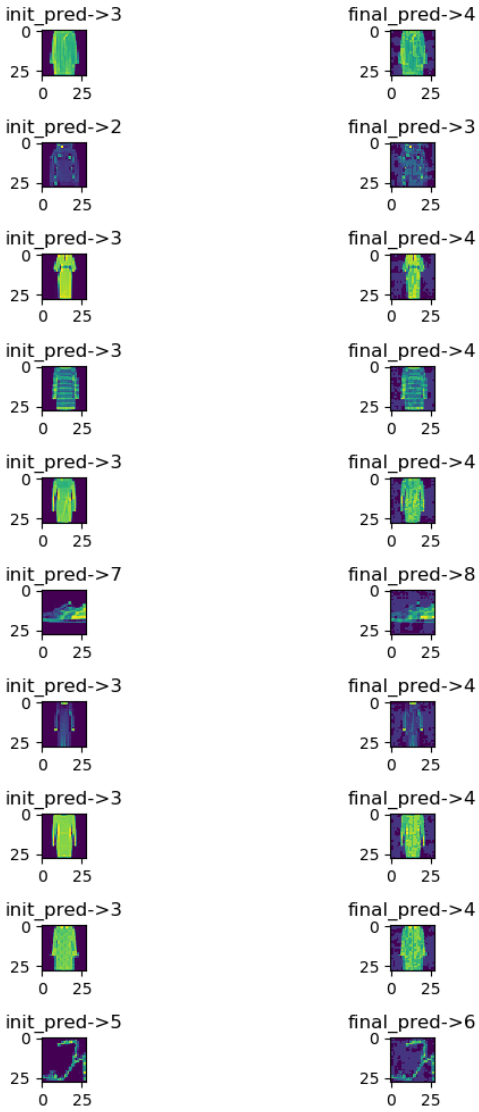

Figure 1: 10 组对抗样本