黑盒攻击报告

信息科学技术学院
智能科学系
彭晗 1801213805

2019 年 3 月 27 日

1 原始模型

原始模型结构未知，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%。【注明：本次实验未完全按照实验要求进行，即并未使用源码提供的 Tensorflow 版本网络模型作为待攻击对象，取而代之的是使用了 pytorch 设计了分类器作为待攻击对象，且此模型与白盒攻击任务中模型一致】

2 对抗攻击

对抗攻击采用样本迁移的方法。首先设计白盒模型，分类器由三层卷积层、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 3*3*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 2*2*20 的卷积核，第四层使用 2*2 核进行 max 池化，第五层卷积层使用 100 个 5*5*50 的卷积核，第六层使用全连接层至 1000 维并使用 relu 作为激活函数，第七层使用全连接层至 10 维并使用 softmax 作为激活函数。

首先筛选出能够在原始黑盒模型中预测正确的样本数据，总计 8989 个。将这些样本分别输入到白盒模型中计算梯度，随后采用对抗算法对样本进行扰动。最后将扰动之后的数据传入黑盒模型，检测是否满足了对抗任务的需求。

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 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 1.54%，选取十组样本进行展示。

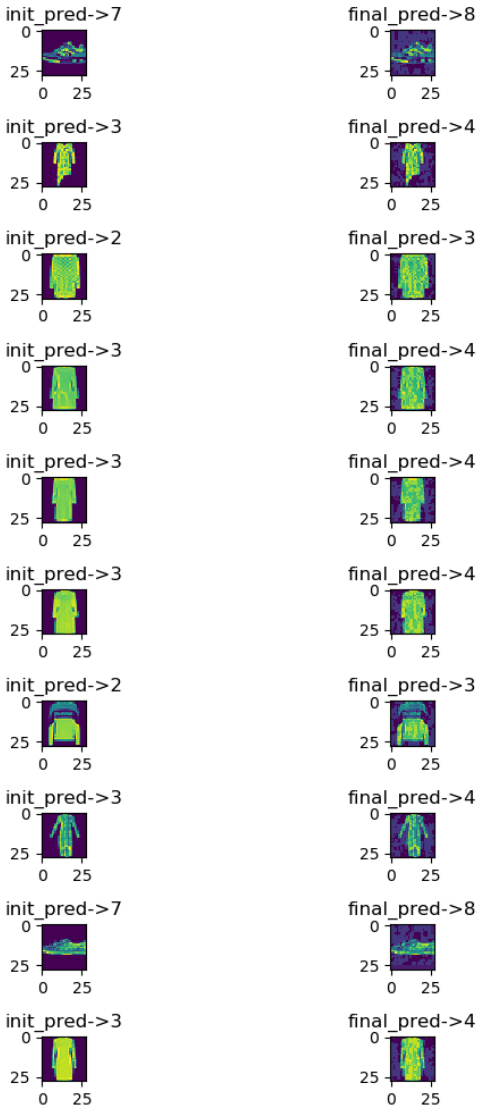

Figure 1: 10 组对抗样本