项目结构

项目分为三部分，完整代码与数据：https://github.com/CarenceLiu/Fashion-MNIST

- 白盒攻击：代码与结果在whiteBoxTask下
  - model.py：自定义的CNN模型与训练过程
  - train_loss.png：训练过程loss变化情况
  - whiteBoxAttack.py：白盒攻击代码，使用对数据的梯度下降
  - result文件夹：中间数据及攻击成功的部分图片对比
- 黑盒攻击：代码及结果在blackBoxTask下
  - model.py：自定义的CNN模型及给定的CNN黑盒模型定义
  - getAttackData.py：选择黑盒攻击的数据（废弃，尝试了用白盒数据攻击黑盒模型，失败）
  - blackBoxAttack.py：黑盒攻击代码，使用MCMC采样，但修改接受条件为预测概率单调增加，迭代2500轮
  - result：中间数据及攻击成功的部分图片对比
- 对抗攻击：代码及结果在AdversaialTask下
  - model.py：自定义CNN模型
  - train_loss.png：训练过程loss变化情况
  - updateModel.py：获得对抗数据并训练新的模型
  - whiteBoxAttack.py：白盒攻击代码，使用对数据的梯度下降
  - blackBoxAttack.py：黑盒攻击代码，使用MCMC采样
  - result：中间数据及攻击成功的部分图片对比

白盒攻击结果

训练集准确度：99.23%

测试集准确度：90.21%（注：给的黑盒模型测试集准确度为89%，不为95%）

白盒攻击成功率：15.9% (159/1000)

黑盒攻击结果

攻击成功率：10.1%（101/1000）

对抗训练结果

新模型训练集准确度：99.07%

新模型测试集准确度：90.30%

旧模型白盒攻击准确度：15.9% 新模型：20.3%

旧模型黑盒攻击准确度：7.70% 新模型：12.70%

经过对抗训练后，模型准确度甚至下降，并没有提升