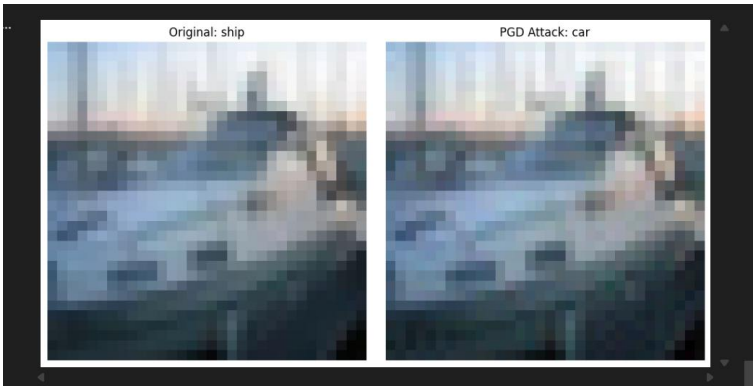
FGSM 复现清单		
复现点 1	对 MNIST 进行 FGSM 攻击	总体成功，错误率偏低
复现点 2	对 Cifar10 进行 FGSM 攻击	总体成功，错误率偏低
复现点 3	对 MNIST 进行对抗训练	鲁棒性提升显著，正则化复现失败
复现点 4	用随机噪声训练做对照实验	总体成功，错误率偏低
复现点 5	对隐藏层进行对抗训练	总体成功

PGD 复现清单		
复现点 1	Cifar10 进行 PGD 攻击	完成
复现点 2	MNIST 的 FGSM 训练模型进行 PGD 攻击	完成
复现点 3	Cifar10 的 PGD 对抗训练（验证 PGD、FGSM）	完成
复现点 4	MNIST 的 PGD 对抗训练（验证 PGD、FGSM）	完成

一、 PGD 攻击

Cifar10

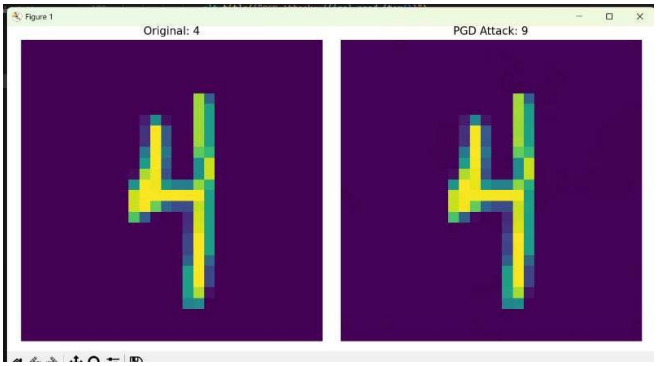


```
正在复现 PGD 攻击，Epsilon = 0.0078，Alpha = 0.0020，Iters = 6...
Files already downloaded and verified
Files already downloaded and verified
Files already downloaded and verified
<ipython-input-1-96cde5547707>:99: FutureWarning: You are using `torch.load` with `weights_only=False`
  model.load_state_dict(torch.load(CHECKPOINT_PATH, map_location=DEVICE))
  0%|          | 3/2000 [00:00<04:41,  7.11it/s]

攻击成功！真值：ship -> 攻击后：car

攻击成功！真值：ship -> 攻击后：car
  0%|          | 6/2000 [00:00<03:11, 10.41it/s]

攻击成功！真值：plane -> 攻击后：ship
100%|██████████| 2000/2000 [00:52<00:00, 38.39it/s]

攻击完成！样本数：2000
攻击成功数 (Adv Success)：1691
攻击成功率 (Error Rate)：84.55%
模型鲁棒准确率 (Robust Acc)：15.45%
```

MNIST

（参数和 cifar10 一样，攻击的是经过 FGSM 对抗训练过的模型）

```
攻击完成！样本数：2000
攻击成功数 (Adv Success)：88
攻击成功率 (Error Rate)：4.40%
模型鲁棒准确率 (Robust Acc)：95.60%
PS C:\Users\MSI\Desktop\MNIST>
```



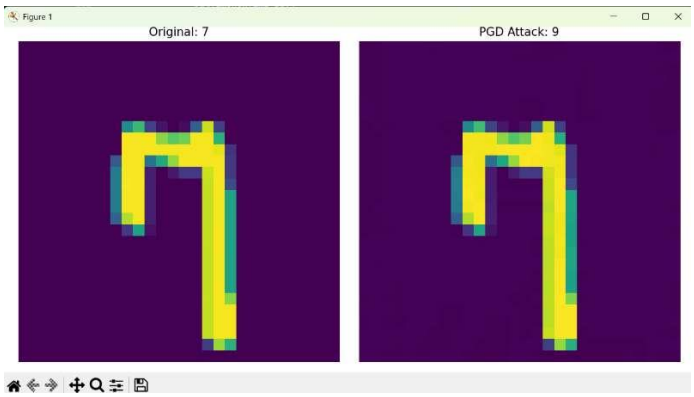


Figure 1
Original: 7
PGD Attack: 9

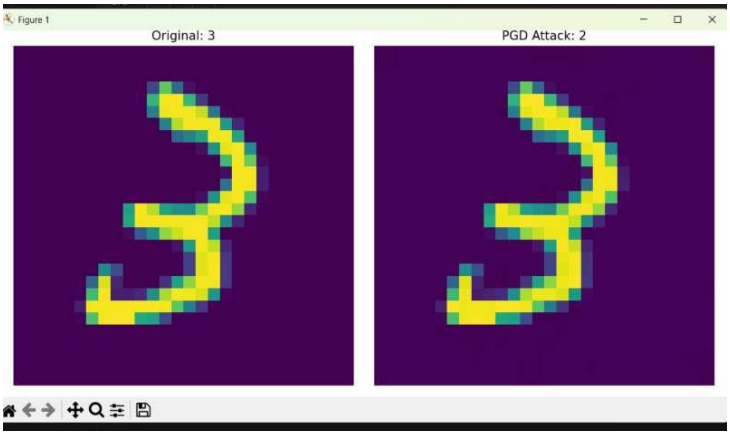


Figure 1
Original: 3
PGD Attack: 2

二、MNIST 的 PGD 对抗训练

PGD 训练 ε = 0.3，iters = 20，α = (ε * 1.5) / iters。

PGD 攻击和训练同参数，FGSM 攻击 ε = 0.3

结果：

第一轮：Results - Clean: 98.39% | FGSM: 46.48% | PGD: 22.57%

（前面做的是一个更弱的攻击，acc4.4%）

最后一轮：Results - Clean: 97.10% | FGSM: 90.52% | PGD: 83.40%

三、cifar10 的 PGD 对抗训练

PGD 训练 ε = 2/255，iters = 6，α = 0.5/255（和前面做的攻击一样的参数）

PGD 攻击 iters = 10（更强的攻击），另外参数一致。FGSM 攻击 ε = 2/255

结果：

第一轮： Clean Acc: 91.17% | FGSM Acc: 69.07% | PGD Acc: 64.78%（前面做出来初始 acc 是 15%）

最后一轮：Clean Acc: 92.39% | FGSM Acc: 76.52% | PGD Acc: 73.91%

四、Trade 训练

在初始模型上续训，结果：PGD acc 49.12%，clean acc 80.80%，我后来又用 FGSM 攻击了一下，FGSM acc 有 76.90%。lr 初始 1e-4，40 轮之后变成 1e-5，第 50 轮早停（patience = 8）