МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт кибербезопасности и цифровых технологий
КБ-4 «Интеллектуальные системы информационной безопасности»

Отчет по лабораторной работе №1

по дисциплине: «Анализ защищенности систем искусственного интеллекта»

Выполнил:
Тимофеев И.О
Группа:
ББМО-01-22

Проверил:
К.т.н. Спирин Андрей Андреевич

Москва, 2023

Шаг 1. Скопировать проект по ссылке в локальную среду выполнения Jupyter (Google Colab) .

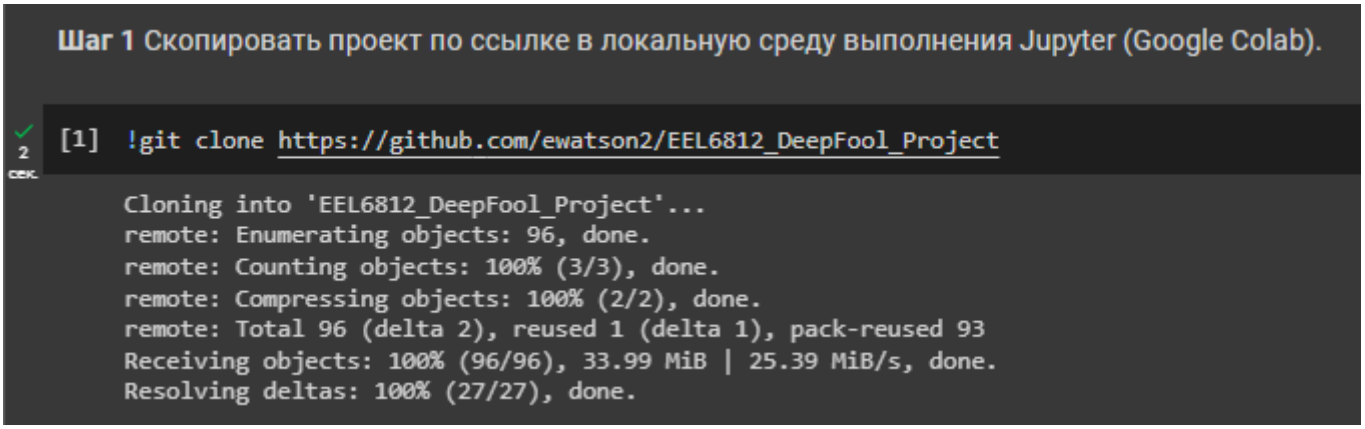

Рисунок 1 – Загрузка проекта

Шаг 2. Сменить директорию исполнения на вновь созданную папку "EEL6812_DeepFool_Project" проекта.

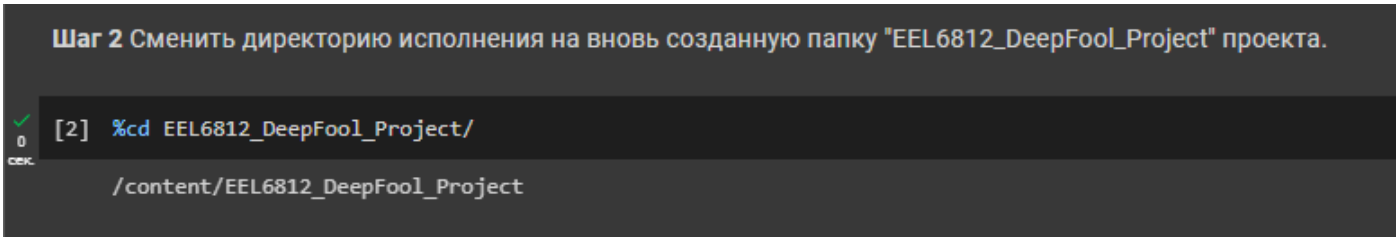

Рисунок 2 – Смена директории

Шаг 3. Выполнить импорт библиотек.

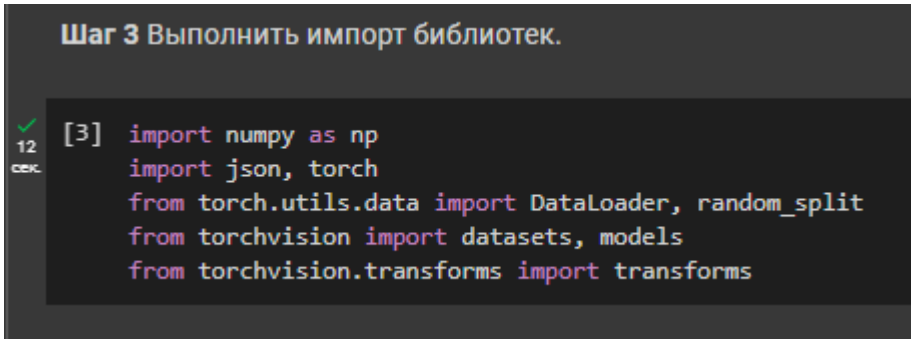

Рисунок 3 – Импорт библиотек

Шаг 4. Выполним импорт вспомогательных библиотек из локальных файлов проекта.

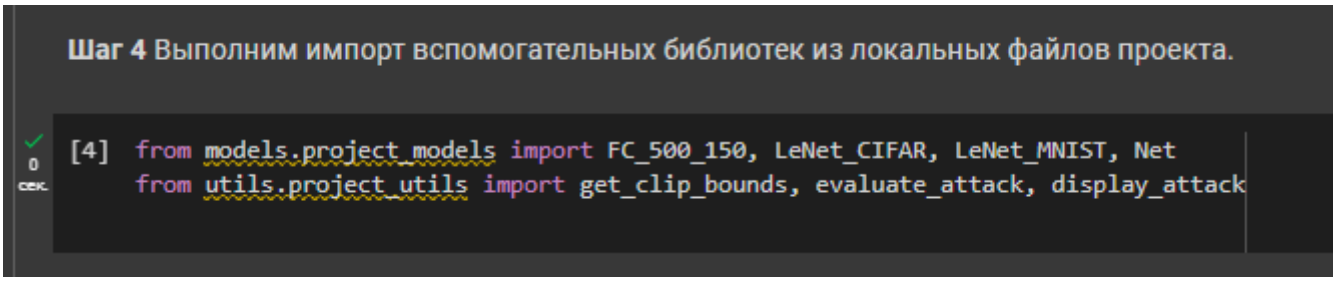

Рисунок 4 – Импорт вспомогательных библиотек

Шаг 5. Установим случайное рандомное значение в виде переменной rand_seed={"Порядковый номер ученика группы в Гугл-таблице"}.

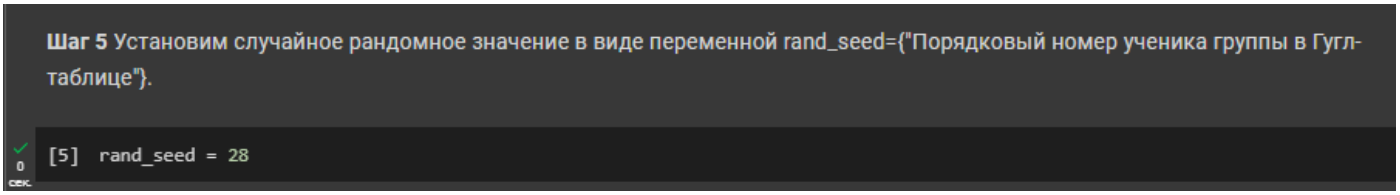

Рисунок 5 – Установка значения для генератора случайных чисел

Шаг 6. Установим указанное значение для np.random.seed и torch.manual_seed.

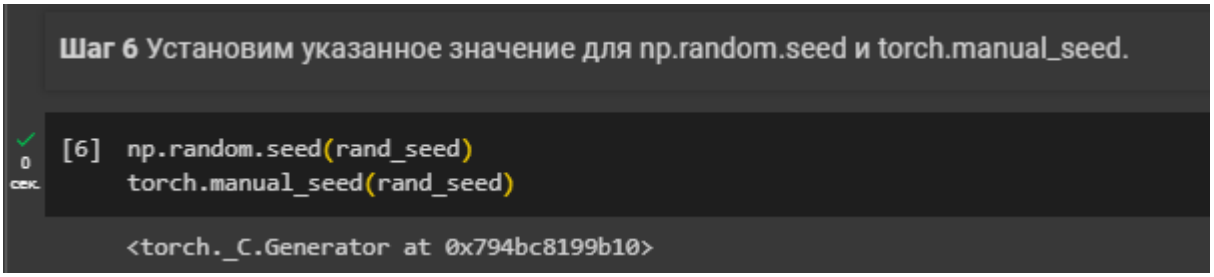

Рисунок 6 – Настройка генераторов случайных чисел

Шаг 7. Используем в качестве устройства видеокарту (Среды выполнения--> Сменить среду выполнения --> T4 GPU).

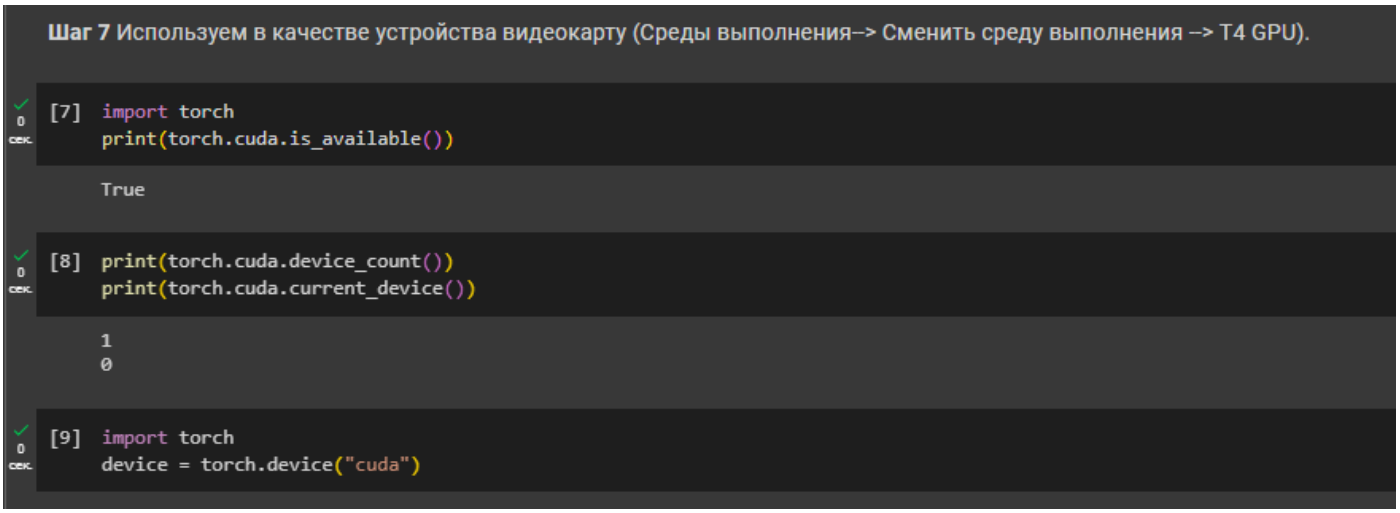

Рисунок 7 – Выбор среды выполнения

Шаг 8. Загрузим датасет MNIST с параметрами mnist_mean = 0.5, mnist_std = 0.5, mnist_dim = 28.

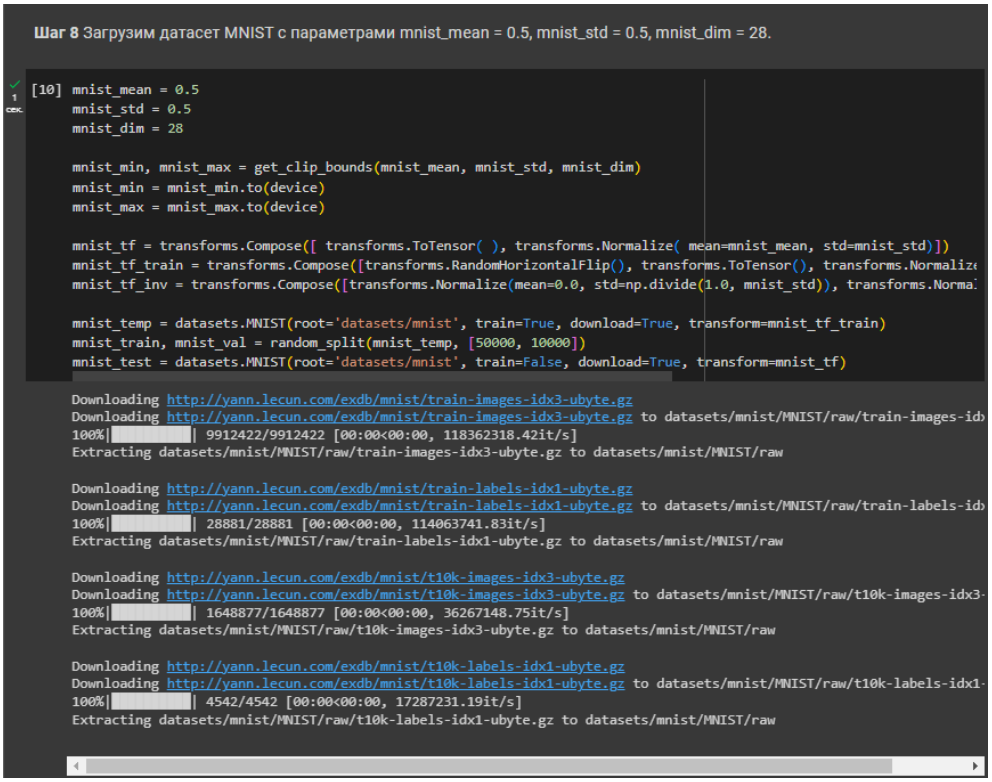

Рисунок 8 – Загрузка датасета MNIST

Шаг 9. Загрузим датасет CIFAR-10 с параметрами.

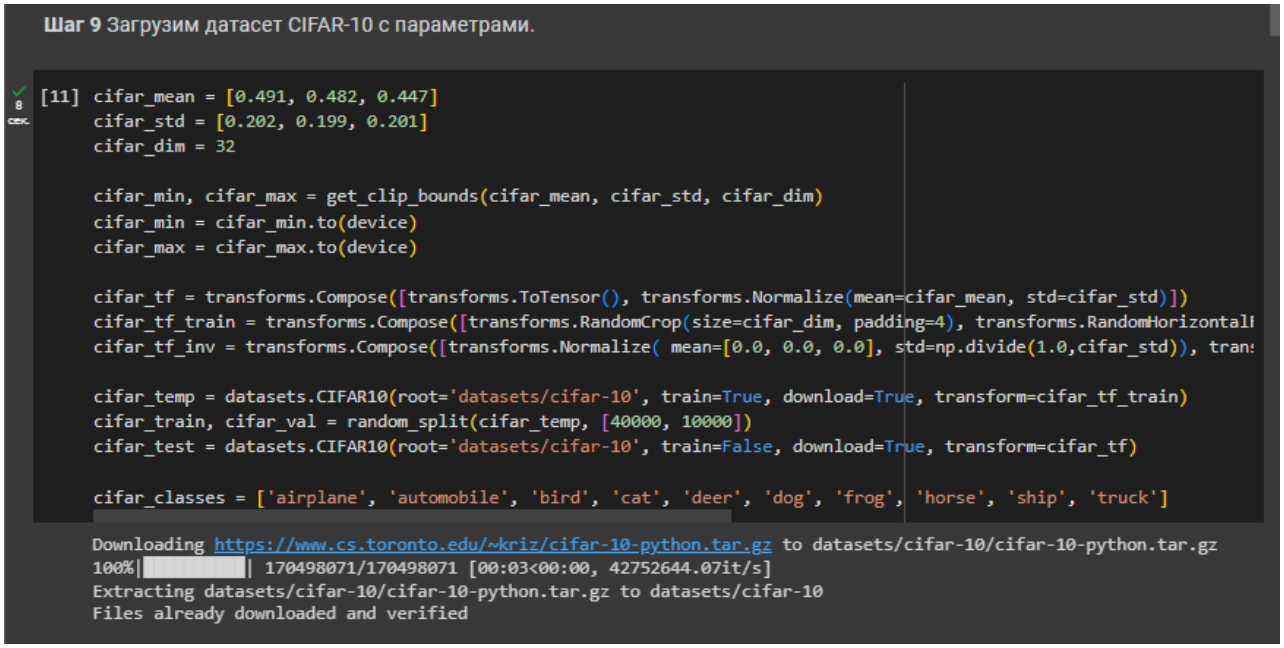

Рисунок 9 – Загрузка датасета CIFAR-10

Шаг 10. Выполним настройку и загрузку DataLoader.

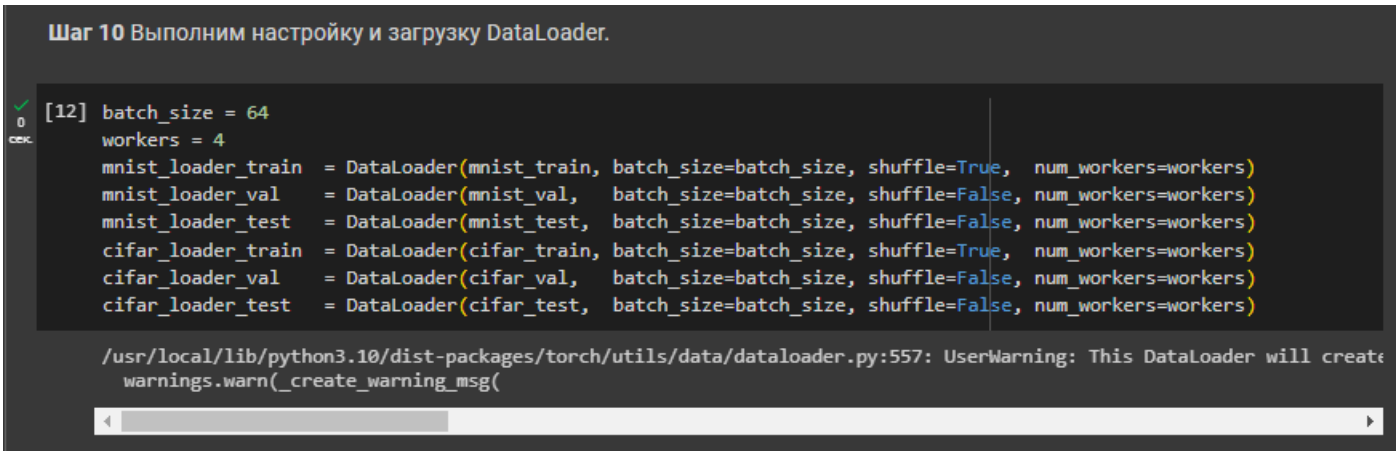

Рисунок 10 – Настройка и загрузка DataLoader

Шаг 11. Загрузим и оценим стойкость модели Network-In-Network Model к FGSM и DeepFool атакам на основе датасета CIFAR-10.

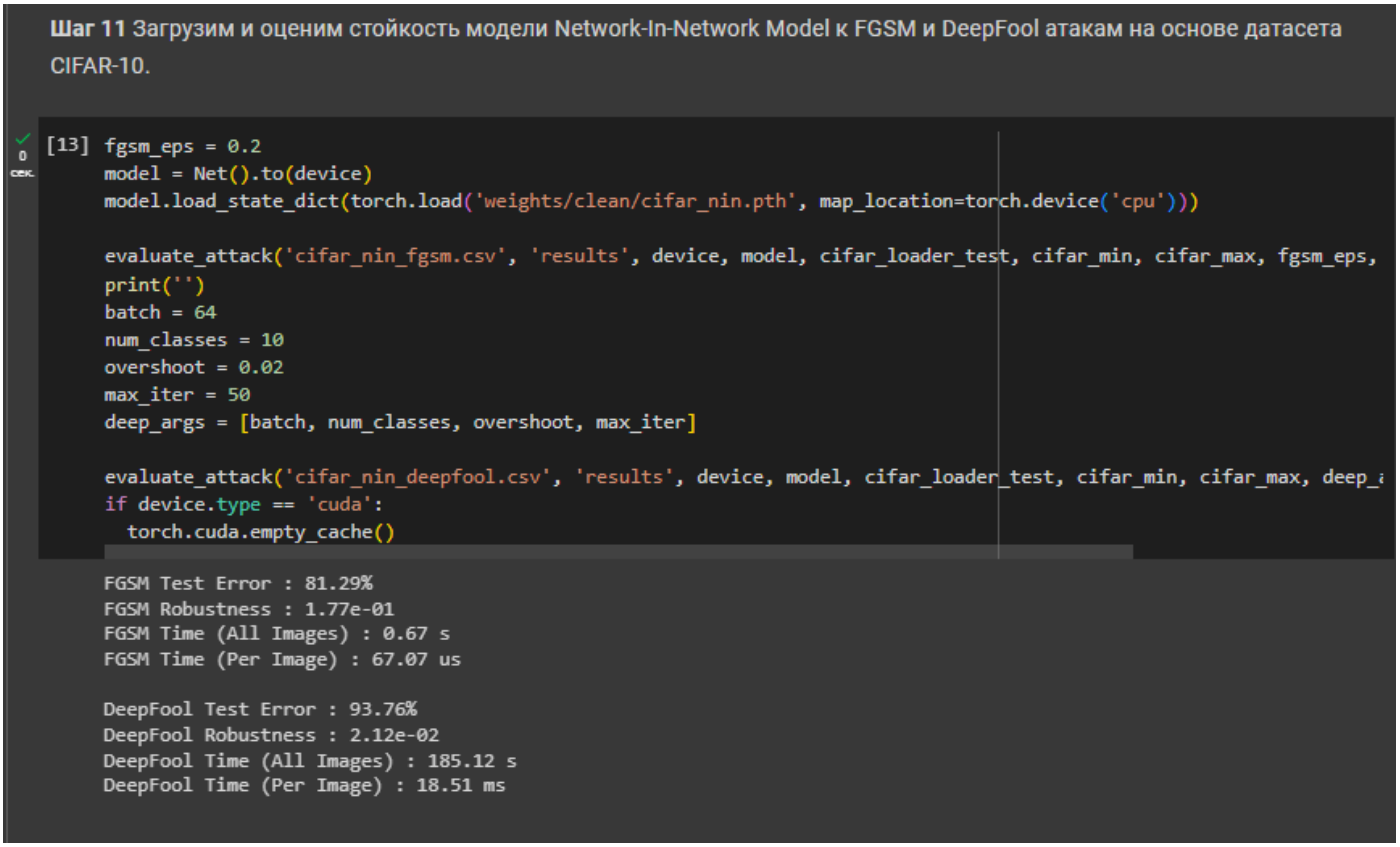

Рисунок 11 – Загрузка и оценка стойкости модели к атакам

Шаг 12. Загрузим и оценим стойкость модели LeNet к FGSM и DeepFool атакам на основе датасета CIFAR-10.

Шаг 12 Загрузим и оценим стойкость модели LeNet к FGSM и DeepFool атакам на основе датасета CIFAR-10.

```
[14] fgsm_eps = 0.6
     model = LeNet_MNIST().to(device)
     model.load_state_dict(torch.load('weights/clean/mnist_lenet.pth', map_location=torch.device('cpu')))

     evaluate_attack('mnist_lenet_fgsm.csv', 'results', device, model, mnist_loader_test, mnist_min, mnist_max, fgsm_eps, is_fgsm=True)

     print('')
     batch = 64
     num_classes = 10
     overshoot = 0.02
     max_iter = 50
     deep_arg = [batch, num_classes, overshoot, max_iter]

     evaluate_attack('mnist_lenet_deepfool.csv', 'results', device, model, mnist_loader_test, mnist_min, mnist_max, deep_arg, is_fgsm=False)
     if device.type == 'cuda':
       torch.cuda.empty_cache()

     FGSM Test Error : 87.89%
     FGSM Robustness : 4.58e-01
     FGSM Time (All Images) : 0.29 s
     FGSM Time (Per Image) : 28.86 us

     DeepFool Test Error : 98.74%
     DeepFool Robustness : 9.64e-02
     DeepFool Time (All Images) : 193.32 s
     DeepFool Time (Per Image) : 19.33 ms
```

Рисунок 12 – Загрузка и оценка стойкости модели к атакам

Шаг 13. Выполним оценку атакующих примеров для сетей.

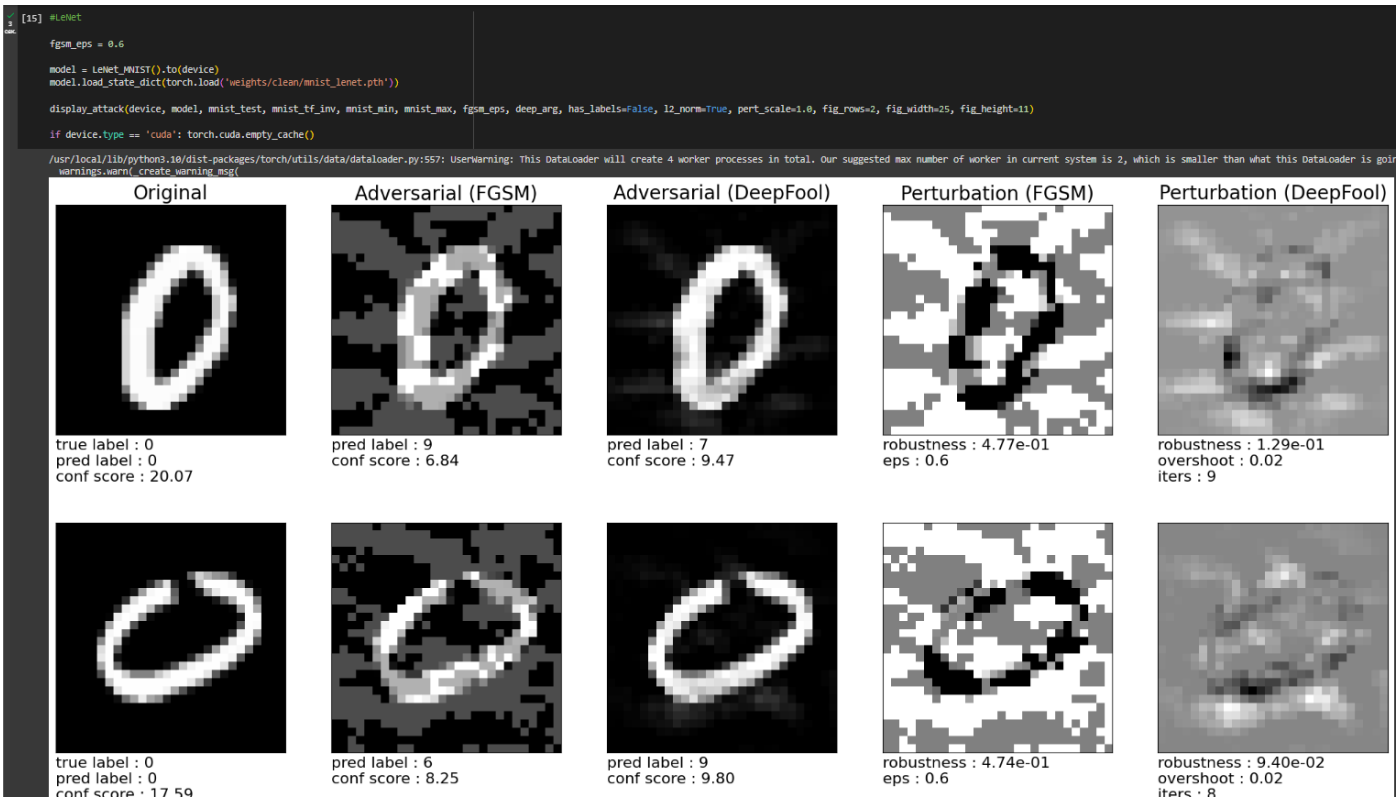

Рисунок 13 – Загрузка модели LeNet и оценка атакующих примеров(MNIST).

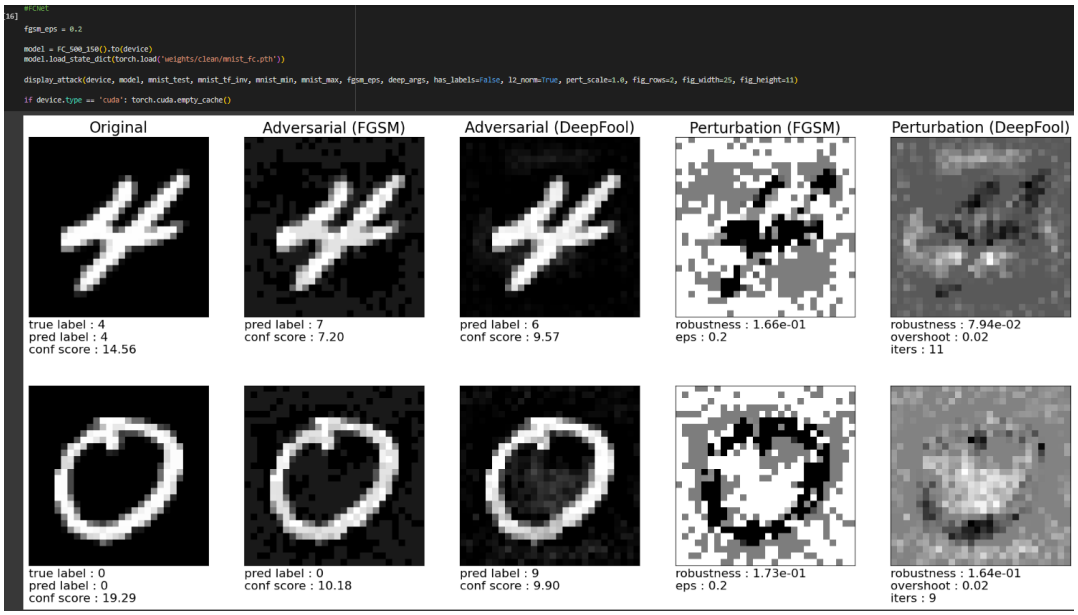

Рисунок 14 – Загрузка модели FC и оценка атакующих примеров(MNIST).

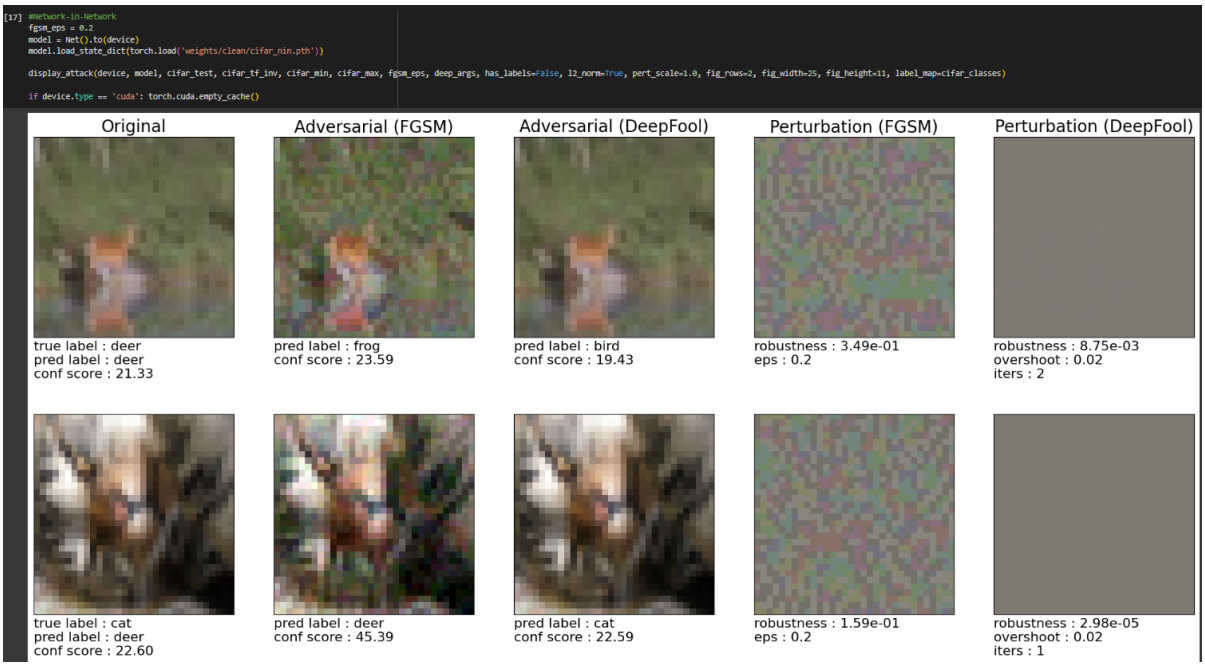

Рисунок 15 – Загрузка модели NiN оценка атакующих примеров (CIFAR-10).

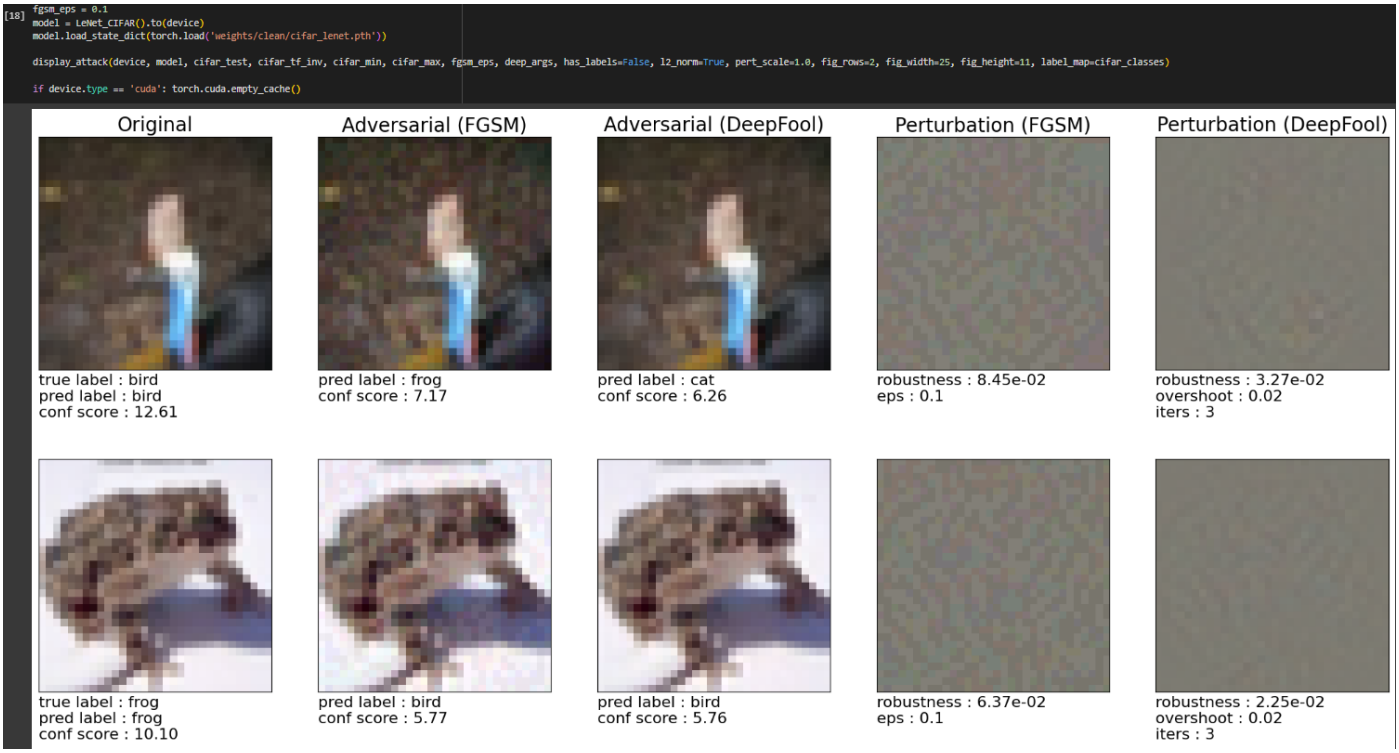

Рисунок 16 – Загрузка модели LeNet оценка атакующих примеров (CIFAR-10)

Шаг 14. Попробуем отразить отличия для fgsm_eps = (0.001, 0.02, 0.5, 0.9, 10) и выявить закономерность/обнаружить отсутсвие влияние параметра eps для сетей FC LeNet на датасете MNIST, NiN LeNEt на датасете CIFAR).

Начнём с оценки FC на MNIST.

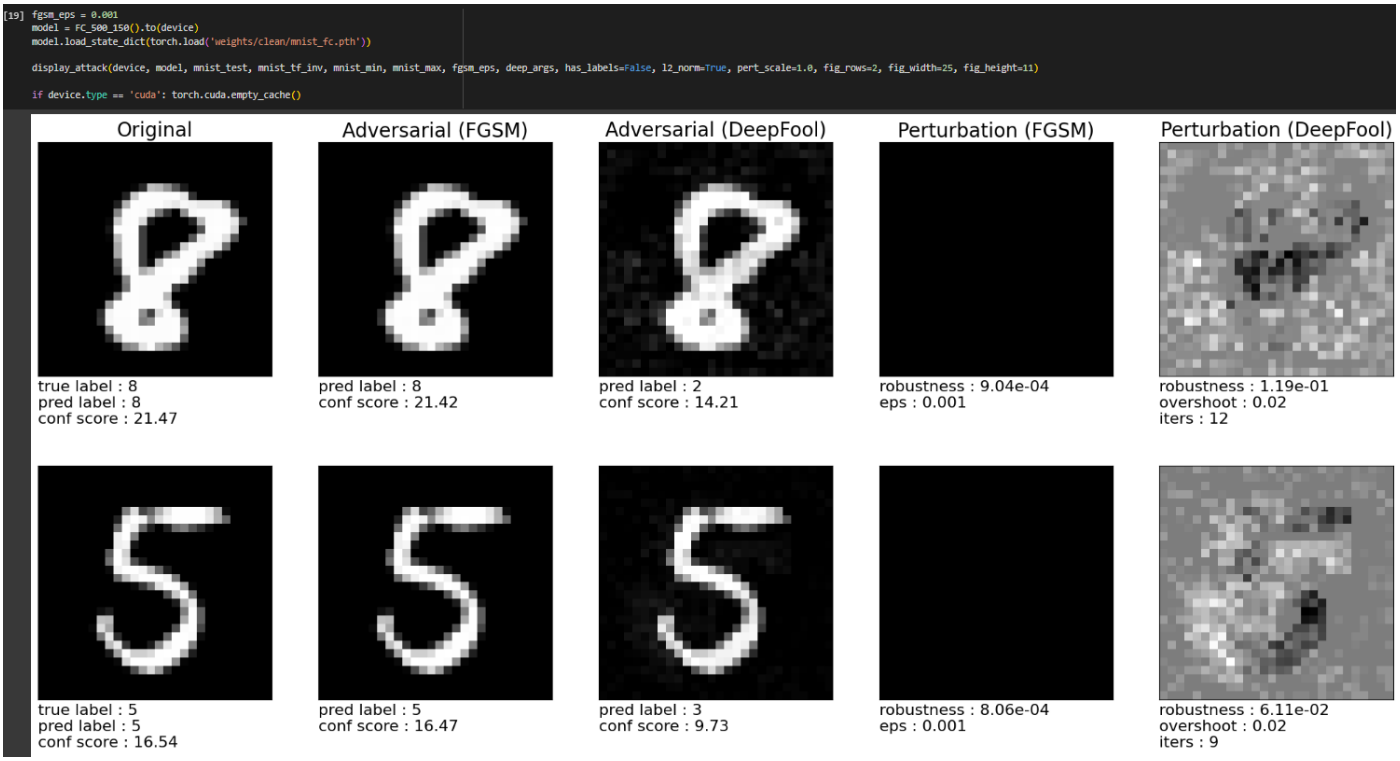

Рисунок 17 – Оценка атакующих примеров для FC на MNIST при eps равном 0.001

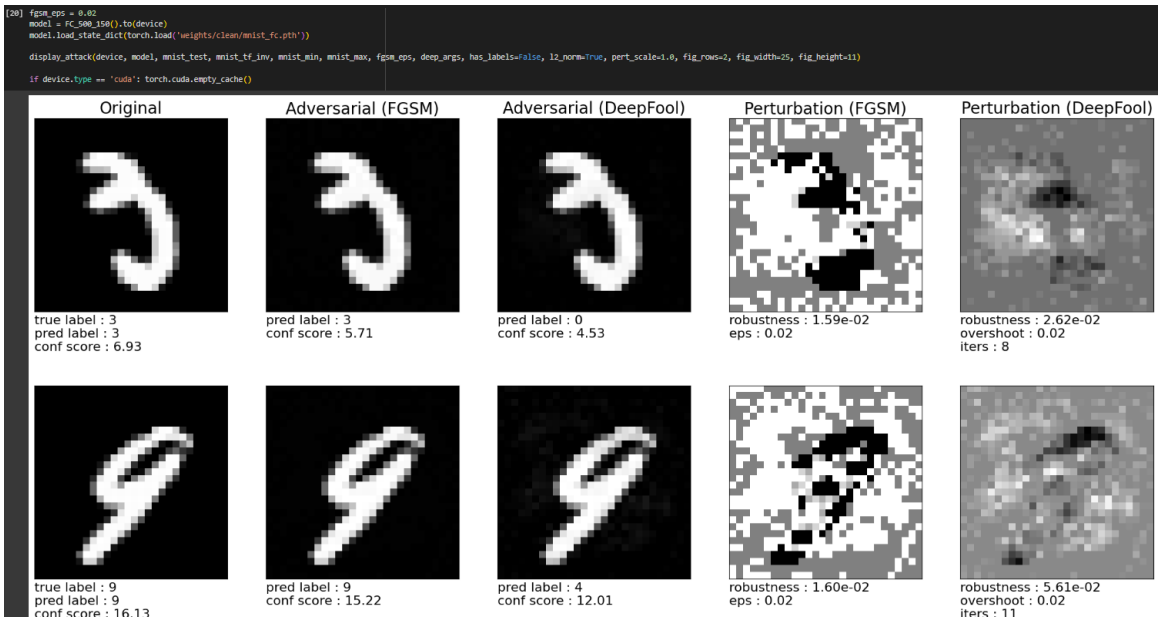

Рисунок 18 – Оценка атакующих примеров для FC на MNIST при eps равном 0.02

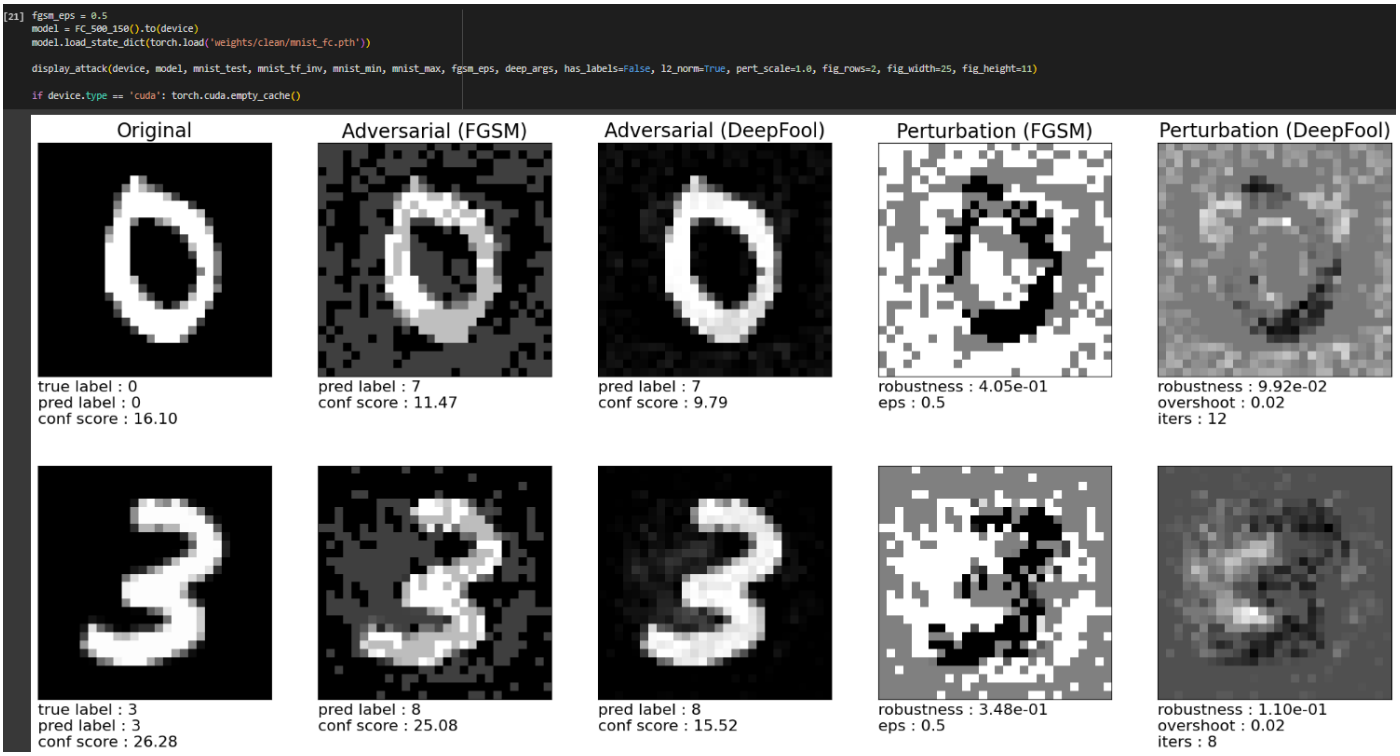

Рисунок 19 – Оценка атакующих примеров для FC на MNIST при eps равном 0.5

Рисунок 20 – Оценка атакующих примеров для FC на MNIST при eps равном 0.9

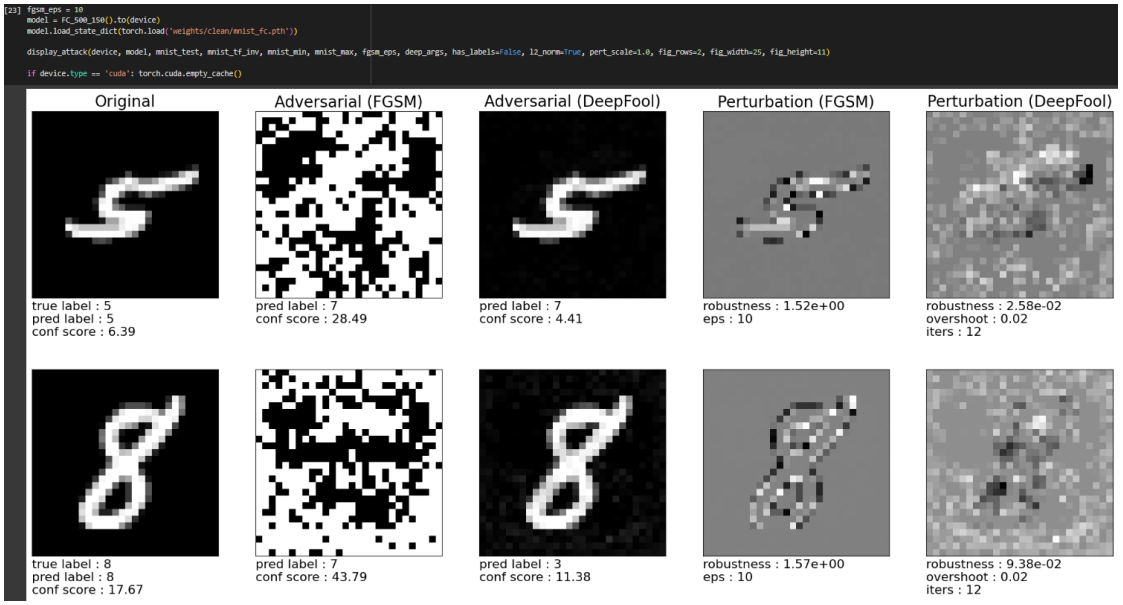

Рисунок 21 – Оценка атакующих примеров для FC на MNIST при eps равном10

Продолжим с оценки LeNet на MNIST.

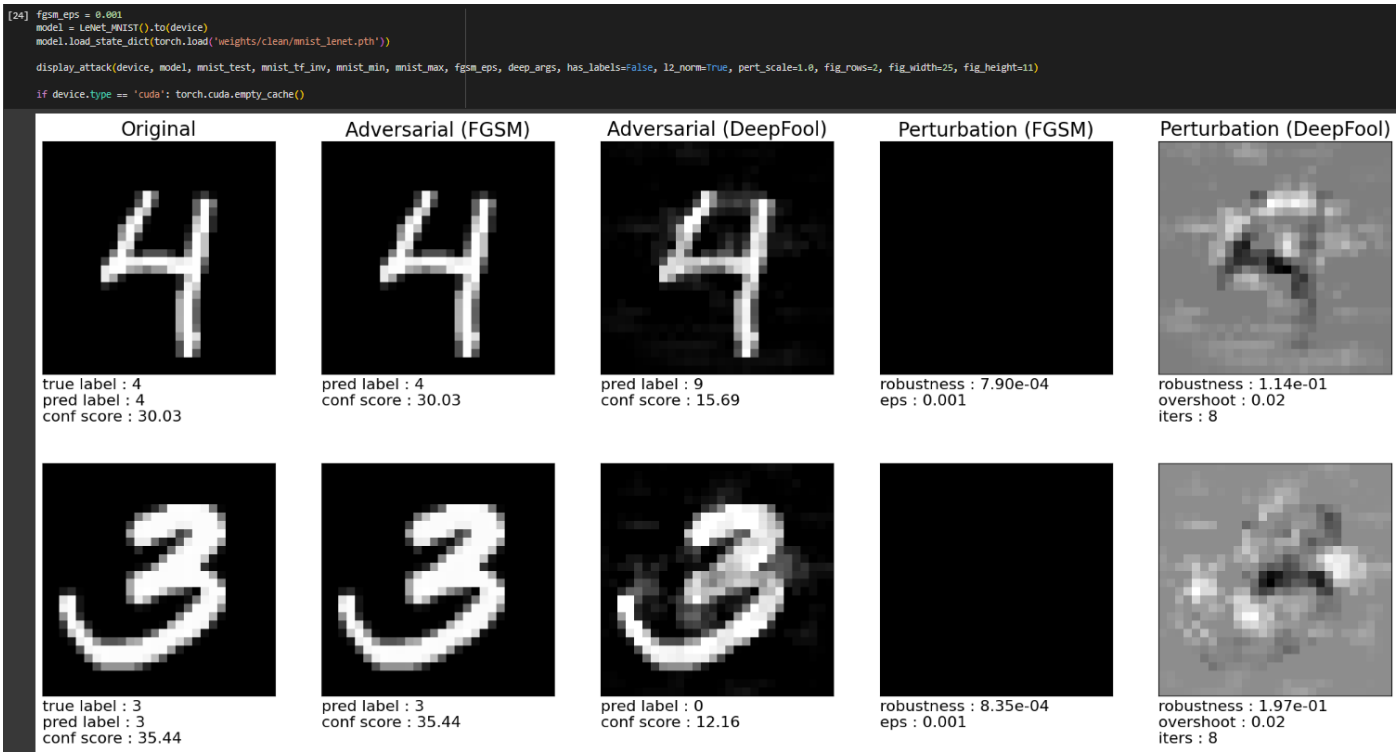

Рисунок 22 – Оценка атакующих примеров для LeNet на MNIST при eps равном0.001

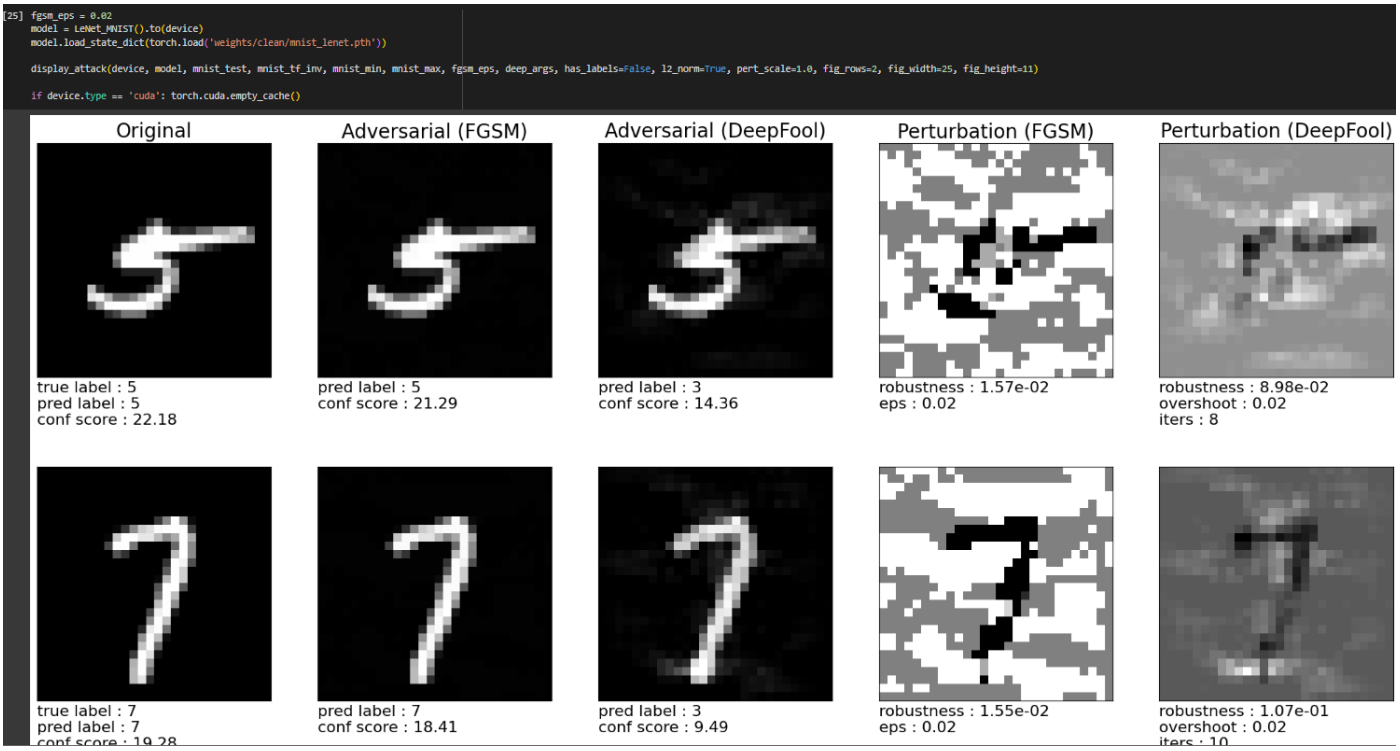

Рисунок 23 – Оценка атакующих примеров для LeNet на MNIST при eps равном 0.02

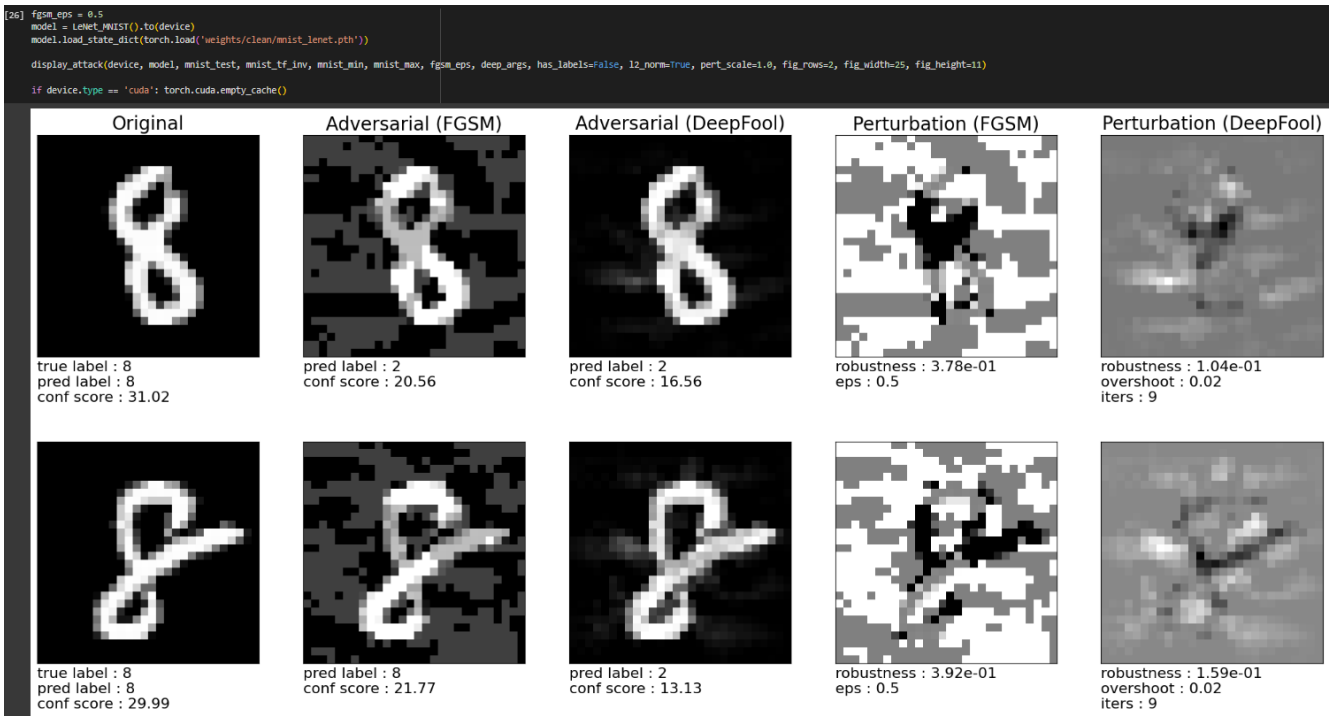

Рисунок 24 – Оценка атакующих примеров для LeNet на MNIST при eps равном 0.5

Рисунок 25 – Оценка атакующих примеров для LeNet на MNIST при eps равном 0.9

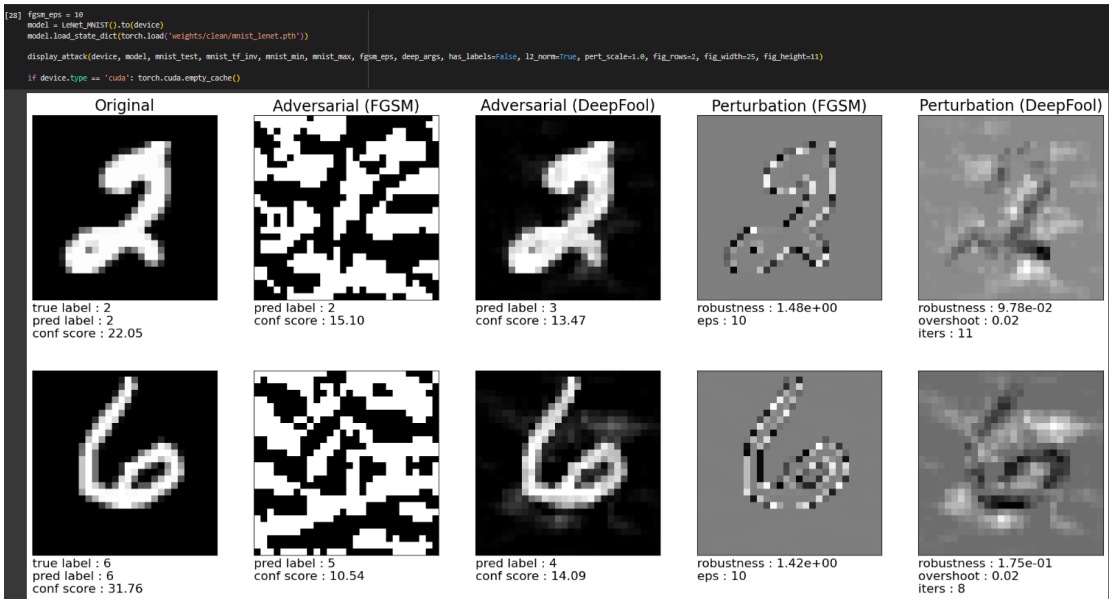

Рисунок 26 – Оценка атакующих примеров для LeNet на MNIST при eps равном 10

Продолжим с оценки NiN на CIFAR-10.

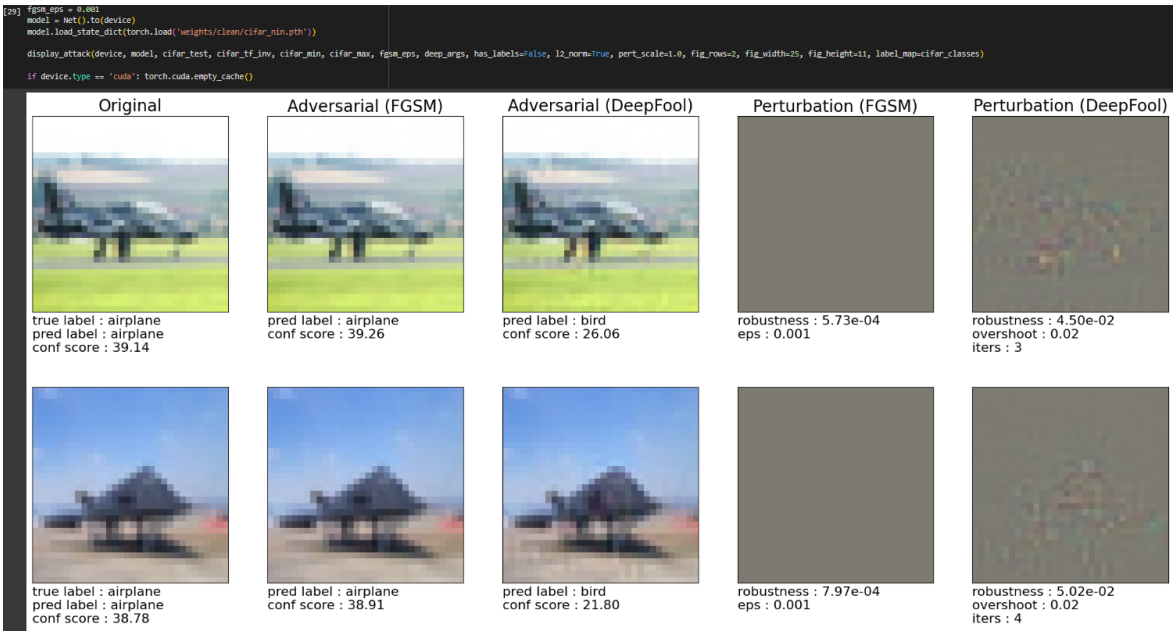

Рисунок 27 – Оценка атакующих примеров для NiN на CIFAR-10 при eps равном 0.001

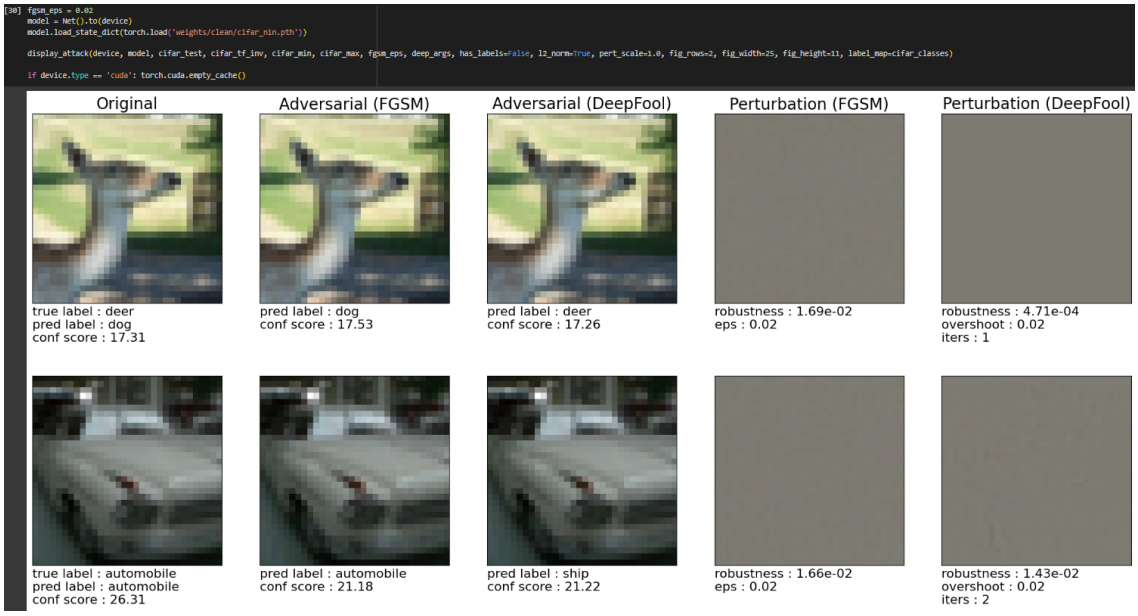

Рисунок 28 – Оценка атакующих примеров для NiN на CIFAR-10 при eps равном 0.02

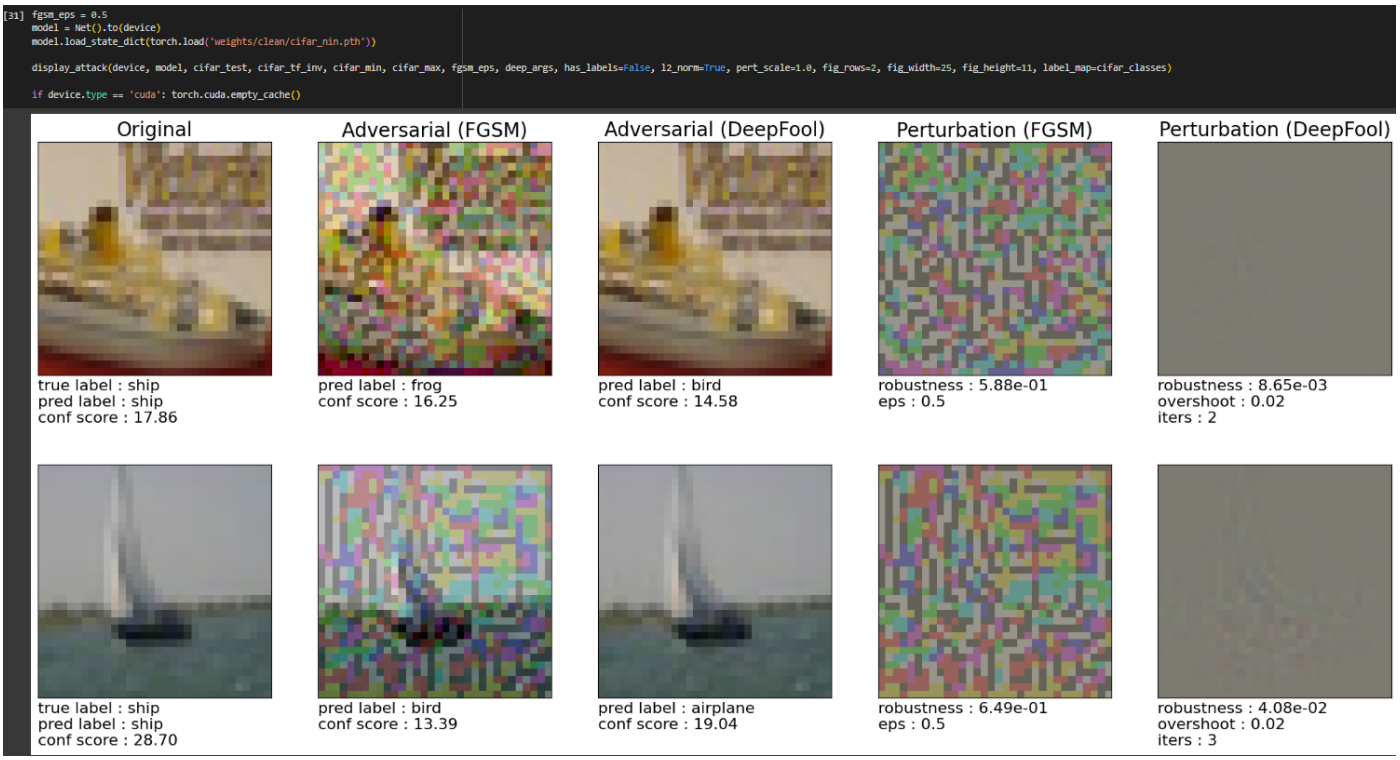

Рисунок 29 – Оценка атакующих примеров для NiN на CIFAR-10 при eps равном 0.5

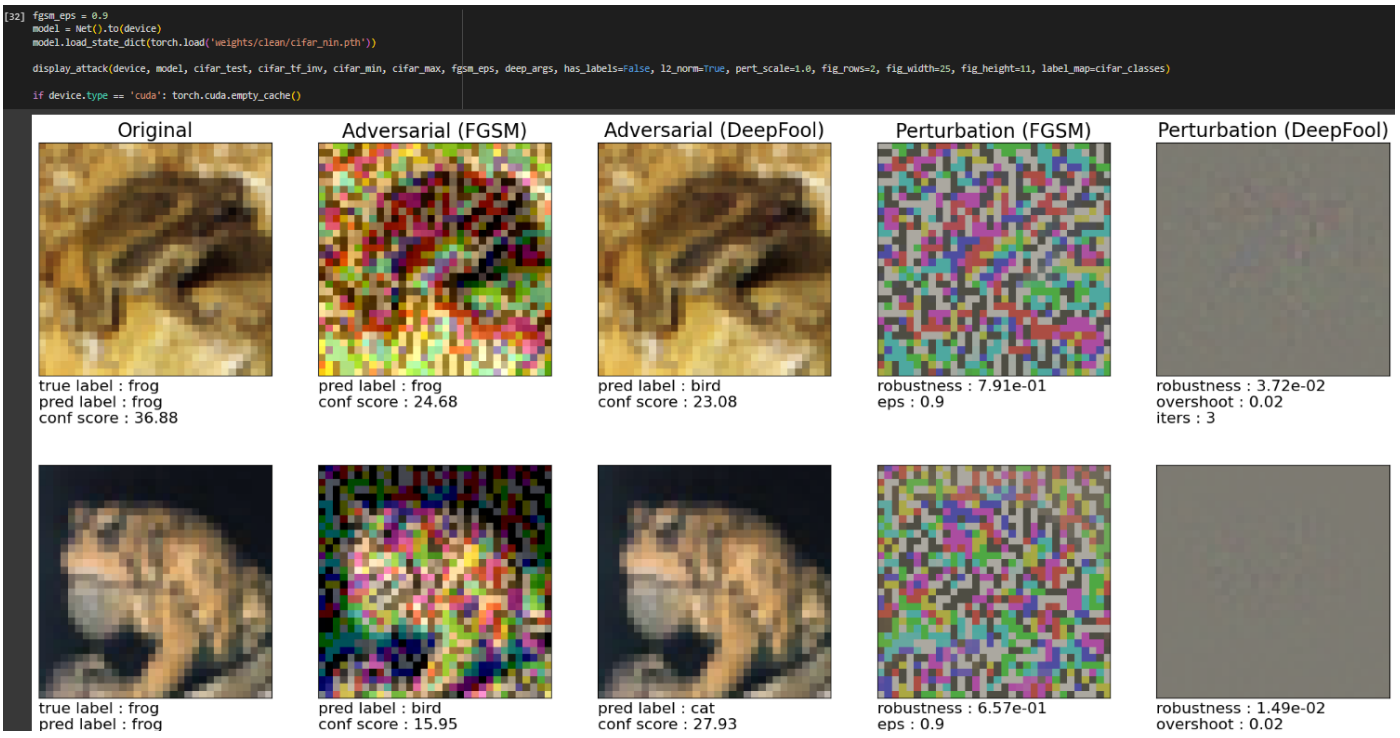

Рисунок 30 – Оценка атакующих примеров для NiN на CIFAR-10 при eps равном 0.9

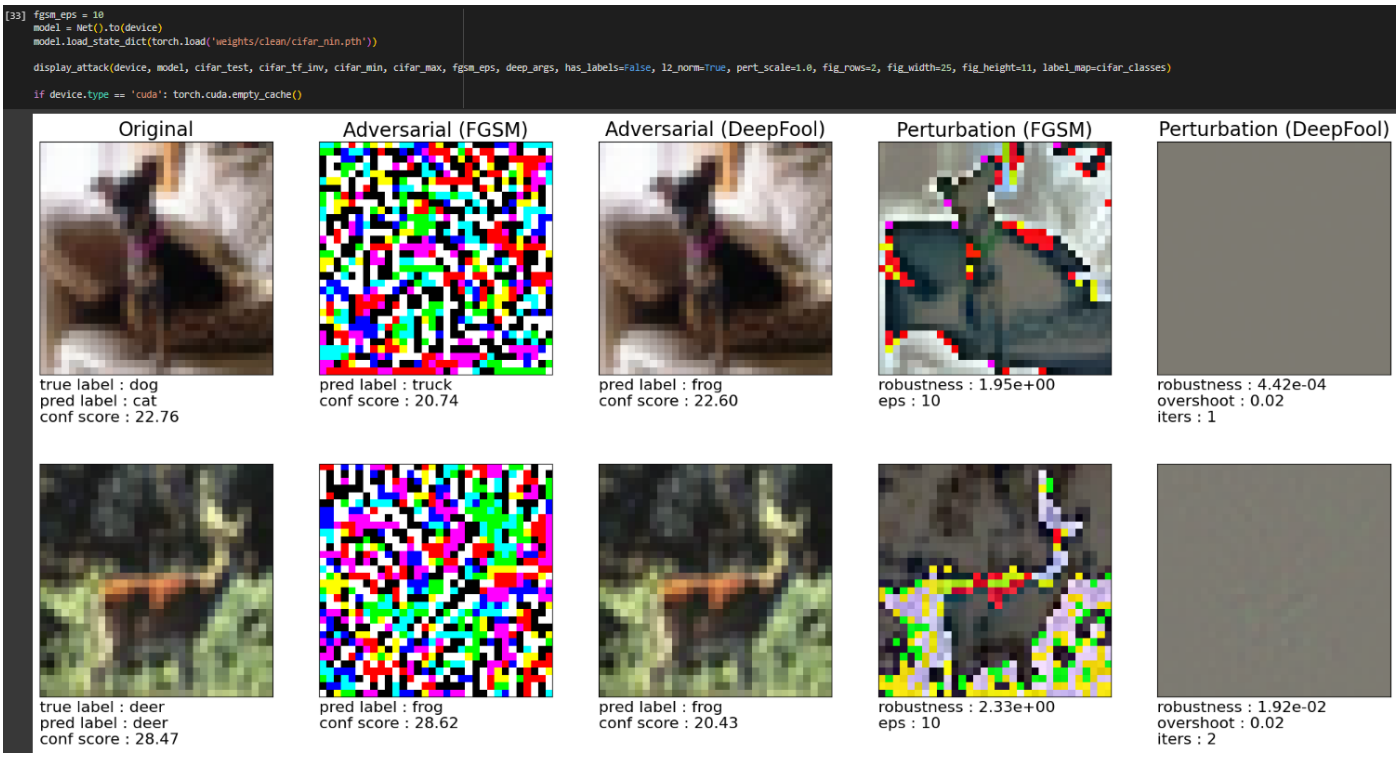

Рисунок 31 – Оценка атакующих примеров для NiN на CIFAR-10 при eps равном 10

Продолжим с оценки LeNet на CIFAR-10.

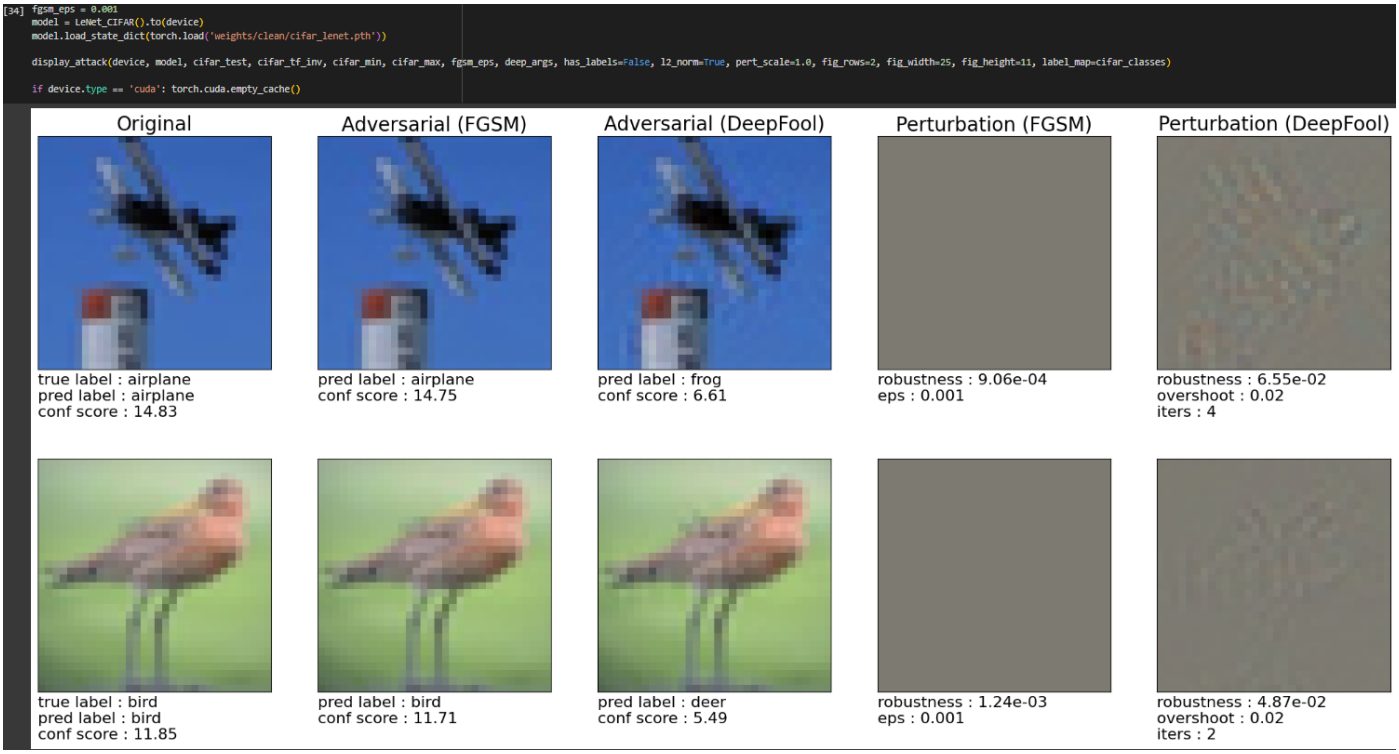

Рисунок 32 – Оценка атакующих примеров для LeNet на CIFAR-10 при eps равном 0.001

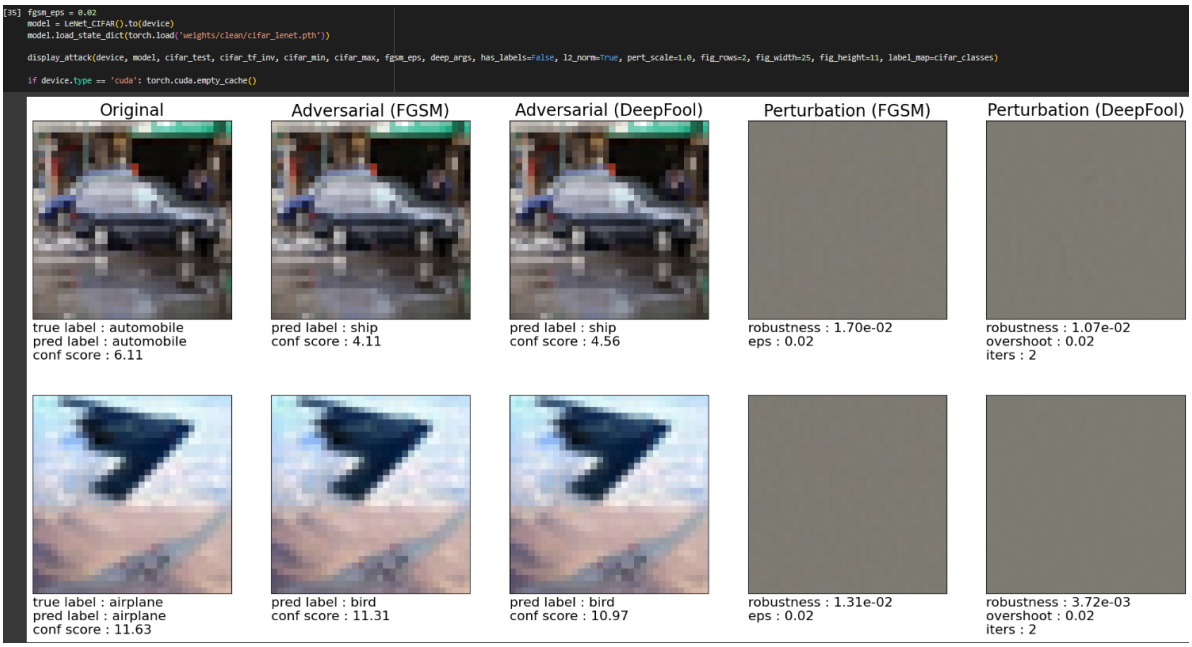

Рисунок 33 – Оценка атакующих примеров для LeNet на CIFAR-10 при eps равном 0.02

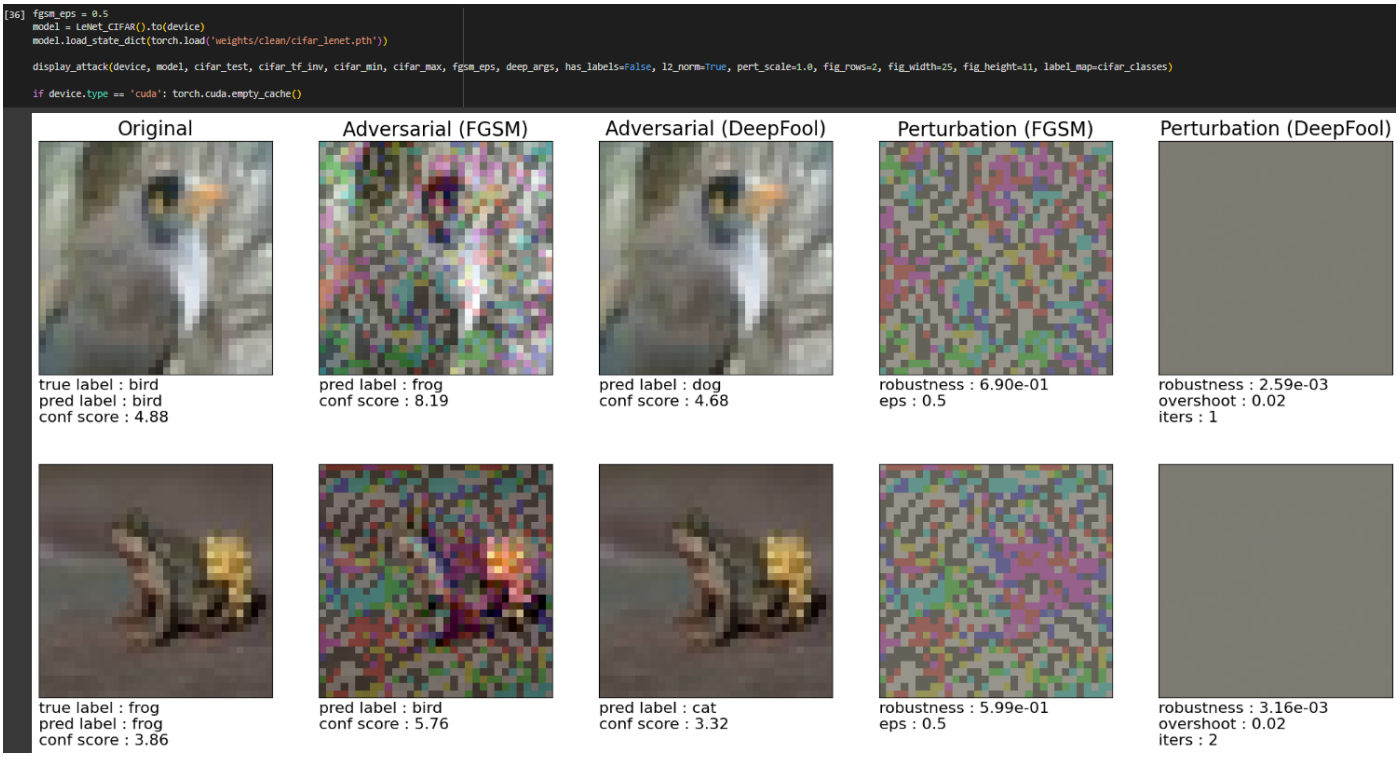

Рисунок 34 – Оценка атакующих примеров для LeNet на CIFAR-10 при eps равном 0.5

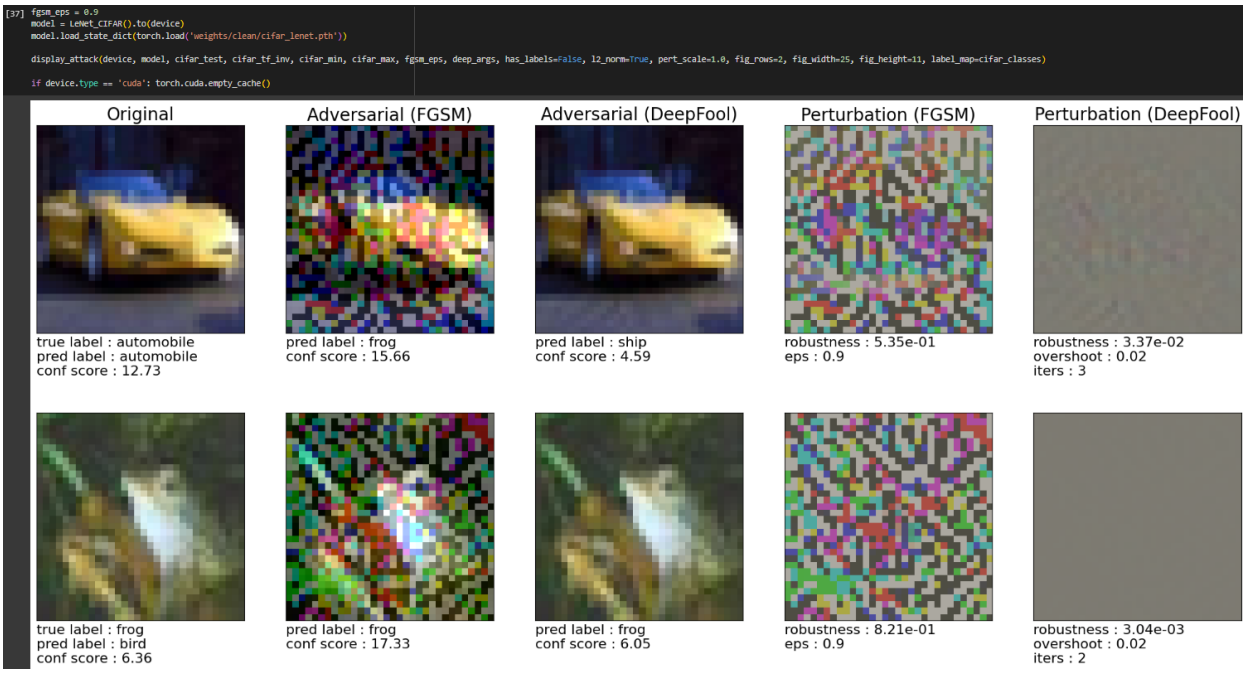

Рисунок 35 – Оценка атакующих примеров для LeNet на CIFAR-10 при epsравном 0.9

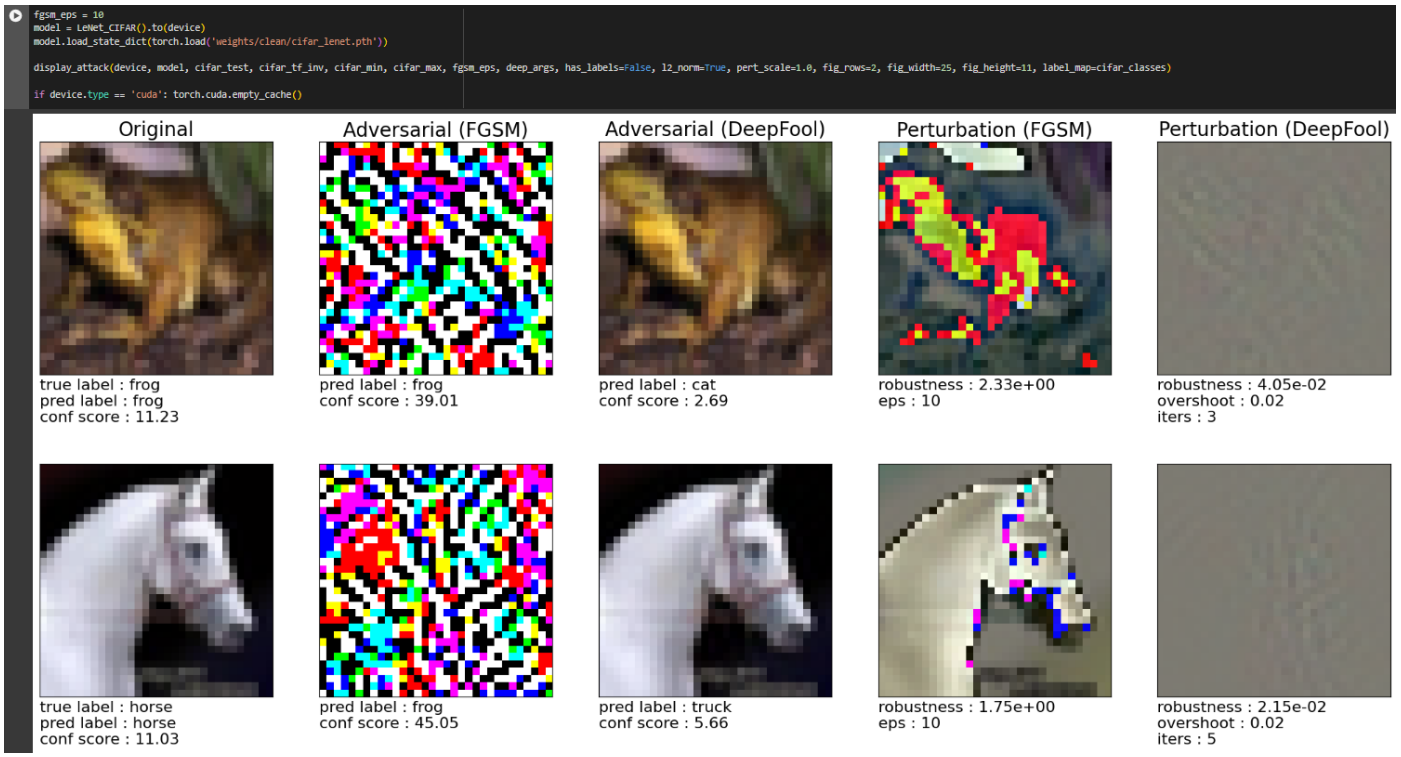

Рисунок 36 – Оценка атакующих примеров для LeNet на CIFAR-10 при eps равном 10

Заключение

В ходе лабораторной работы мы исследовали влияние различных значений eps на модели при атаках. Было обнаружено, что увеличение eps в FGSM приводит к искажению изображения, которое становится заметным для человека. Однако, слишком низкое значение eps может не вызвать ошибку в классификации модели. В случае с DeepFool, изображение практически не изменяется при любом значении eps, но модель всегда ошибается.