|  |  |  |
| --- | --- | --- |
|  |  |  |
| МИНОБРНАУКИ РОССИИ | | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | |
| Институт кибербезопасности и цифровых технологий  Кафедра КБ-4 «Интеллектуальные системы информационной безопасности» | | |

**Отчёт по лабораторной работе №1**

**по дисциплине «Анализ защищенности систем искусственного интеллекта»**

Выполнил:

Евдокимов А.М.

Группа: ББМО-02-23

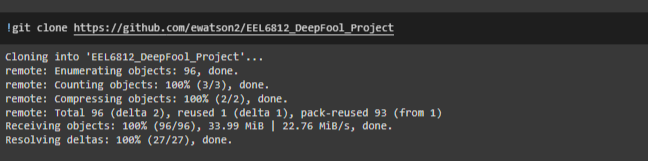
Москва - 2024

**Цель лабораторной работы**

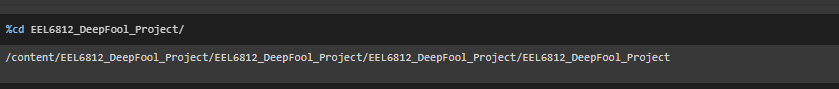
В данной лабораторной работе необходимо выявить закономерность или обнаружить отсутствие влияния параметра fgsm\_eps на стойкость моделей к атаке. Закономерности или их отсутствие необходимо выявить для сети FC LeNet на датасете MNIST и для сети NiN LeNEt на датасете CIFAR.

**Выполнение лабораторной работы**

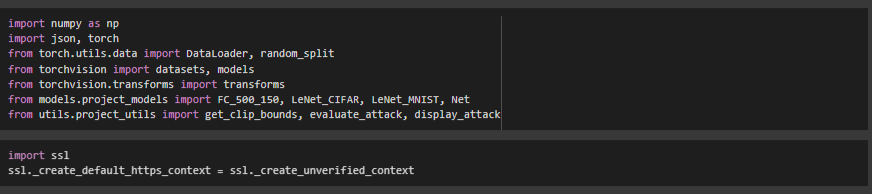
1. Скопировать проект по ссылке в локальную среду выполнения Jupyter (Google Colab) https://github.com/ewatson2/EEL6812\_DeepFool\_Project



2. Сменить директорию исполнения на вновь созданную папку "EEL6812\_DeepFool\_Project" проекта.



3. Выполнить импорт библиотек:

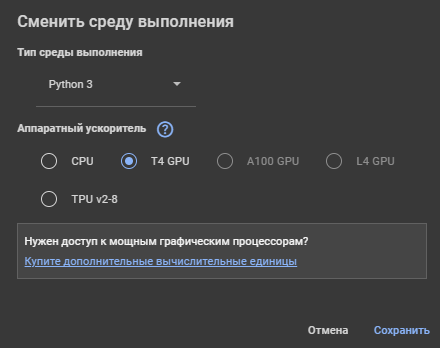


4. Выполнить импорт вспомогательных библиотек из локальных файлов проекта

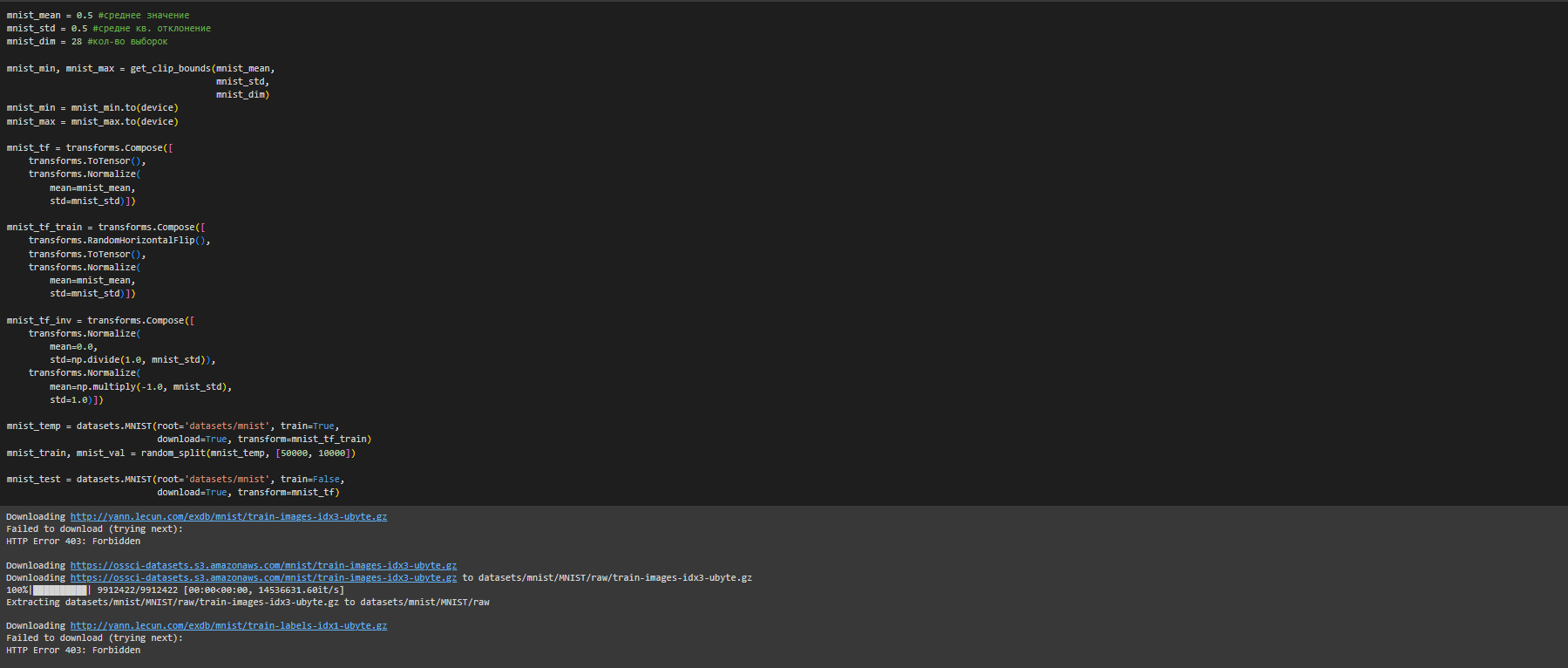
5. Установить значение в виде переменной rand\_seed. Установить указанное значение для np.random.seed и torch.manual\_seed.



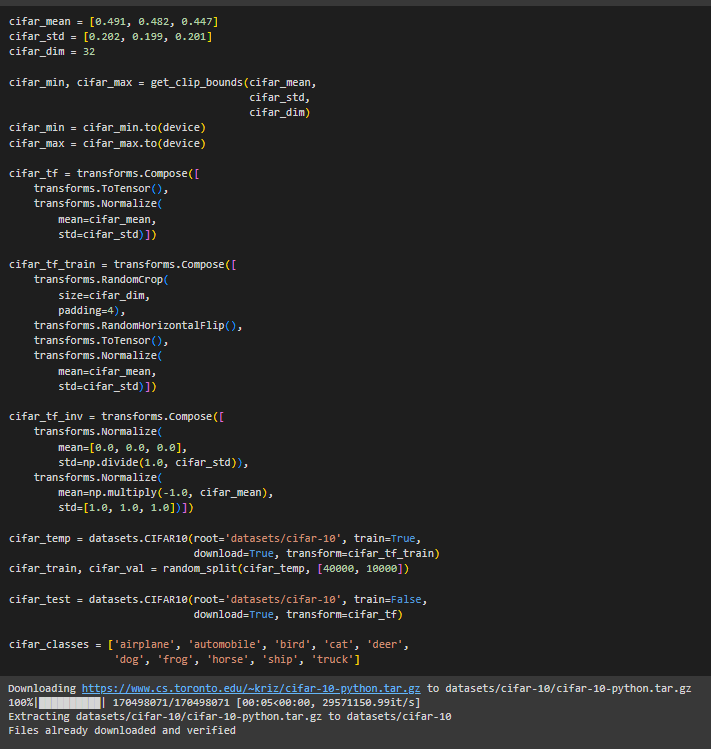
6. Использовать в качестсве устройства видеокарту (Среды выполнения--> Сменить среду выполнения --> T4 GPU)



7. Загрузить датасет MNIST c параметрами mnist\_mean = 0.5, mnist\_std = 0.5, mnist\_dim = 28

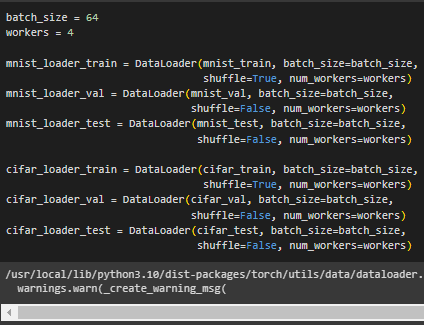


8. Загрузить датасет CIFAR-10 c параметрами cifar\_mean = [0.491, 0.482, 0.447] cifar\_std = [0.202, 0.199, 0.201] cifar\_dim = 32

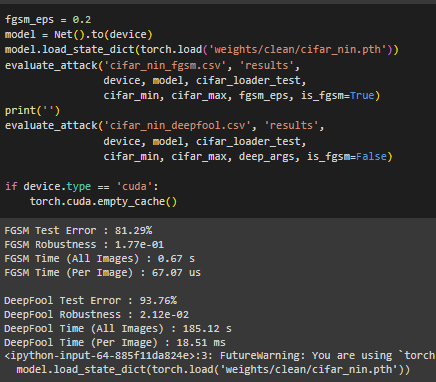


9. Выполнить настройку и загрузку

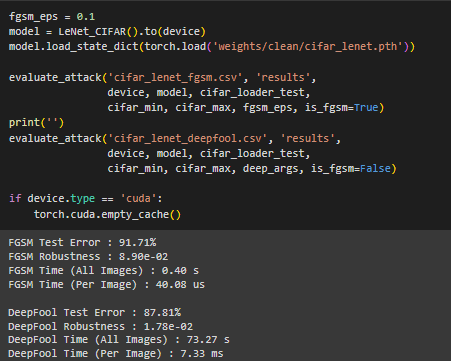
DataLoader batch\_size = 64 workers = 4



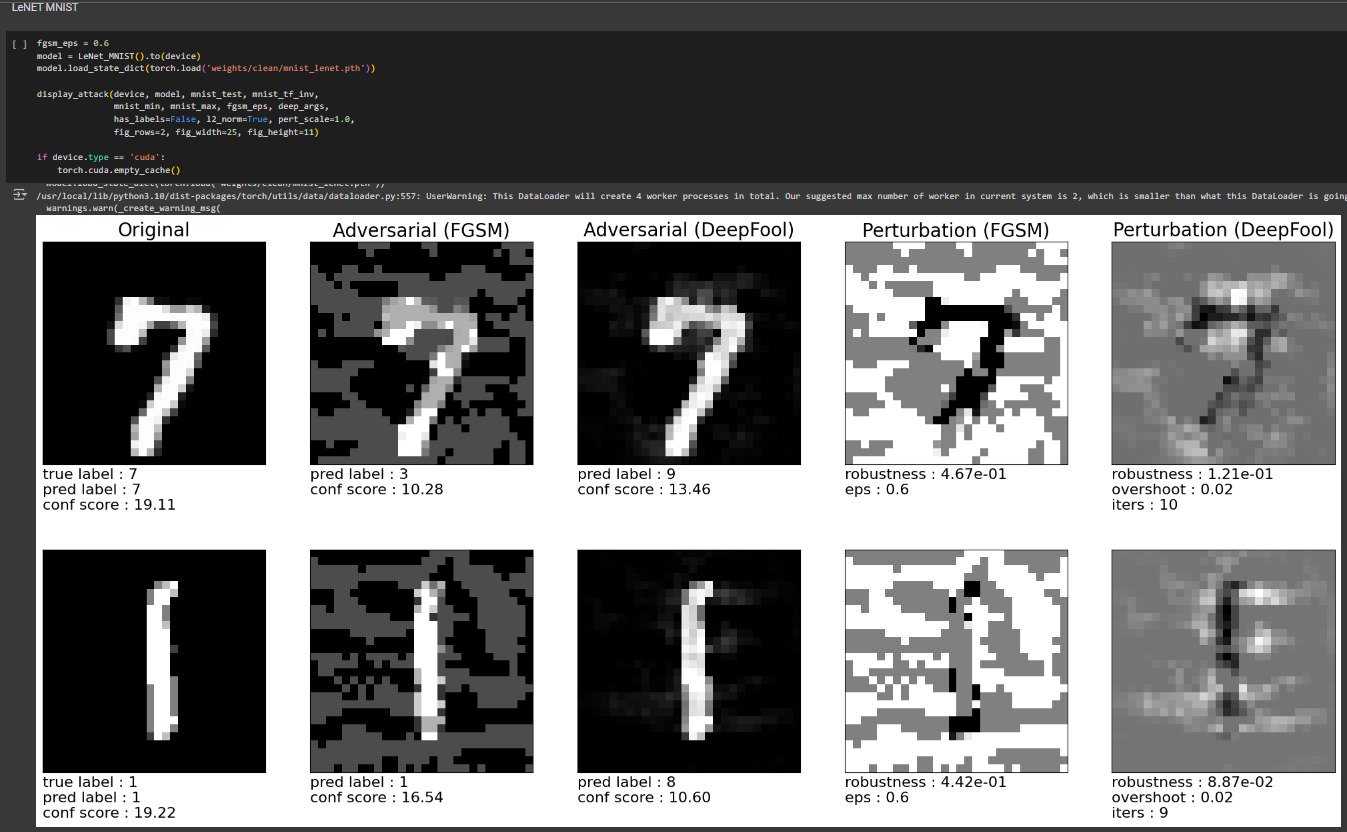
10. Загрузить и оценить стойкость модели Network-In-Network Model к FGSM и DeepFool атакам на основе датасета CIFAR-10

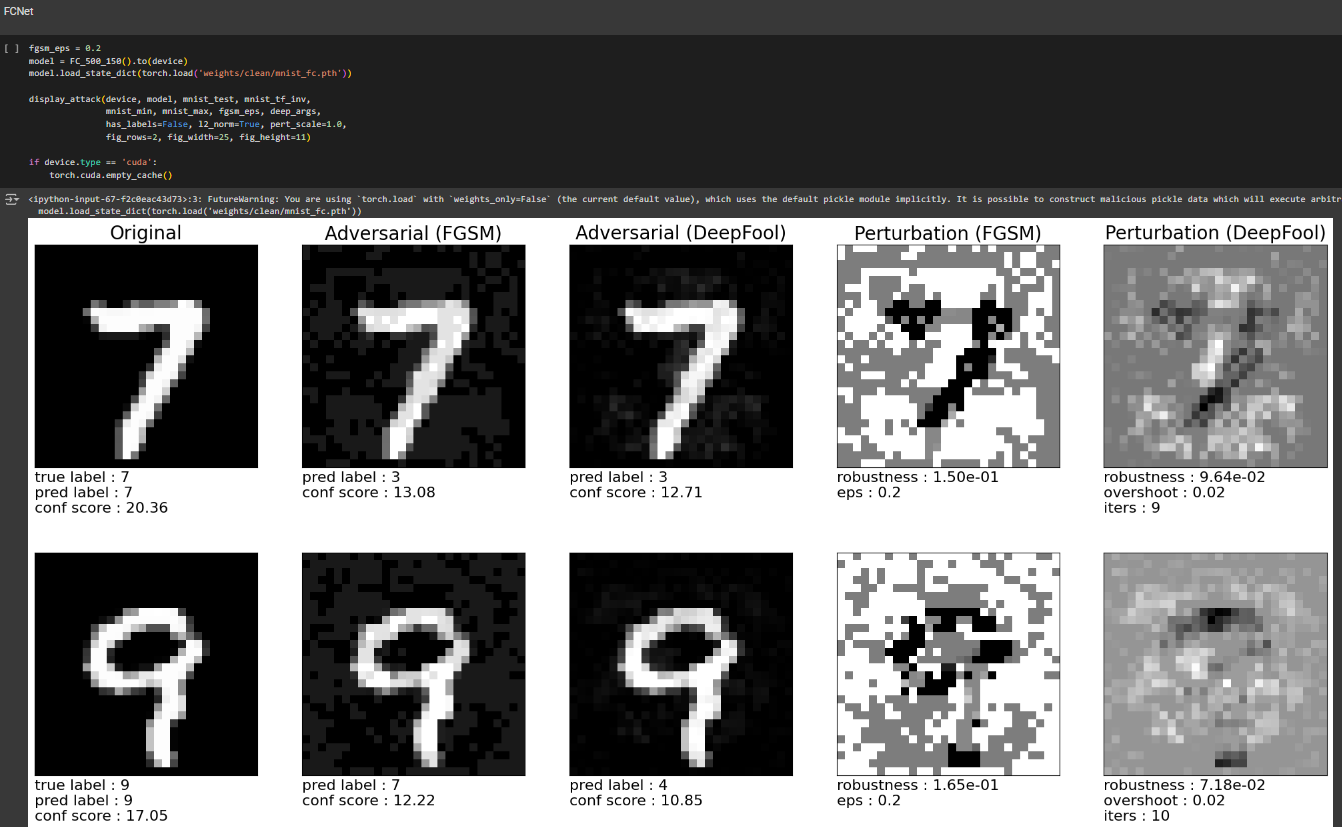


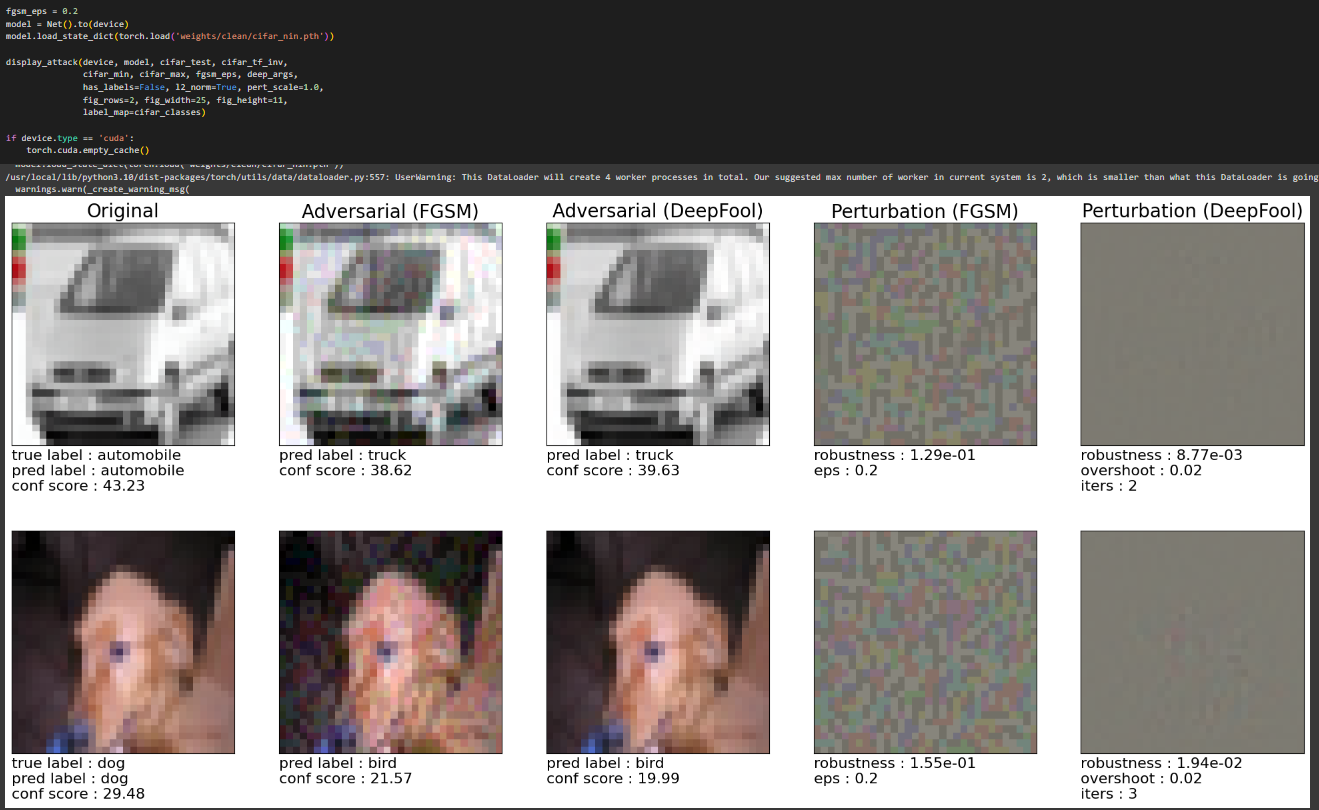
11. Загрузить и оценить стойкость модели LeNet к FGSM и DeepFool атакам на основе датасета CIFAR-10

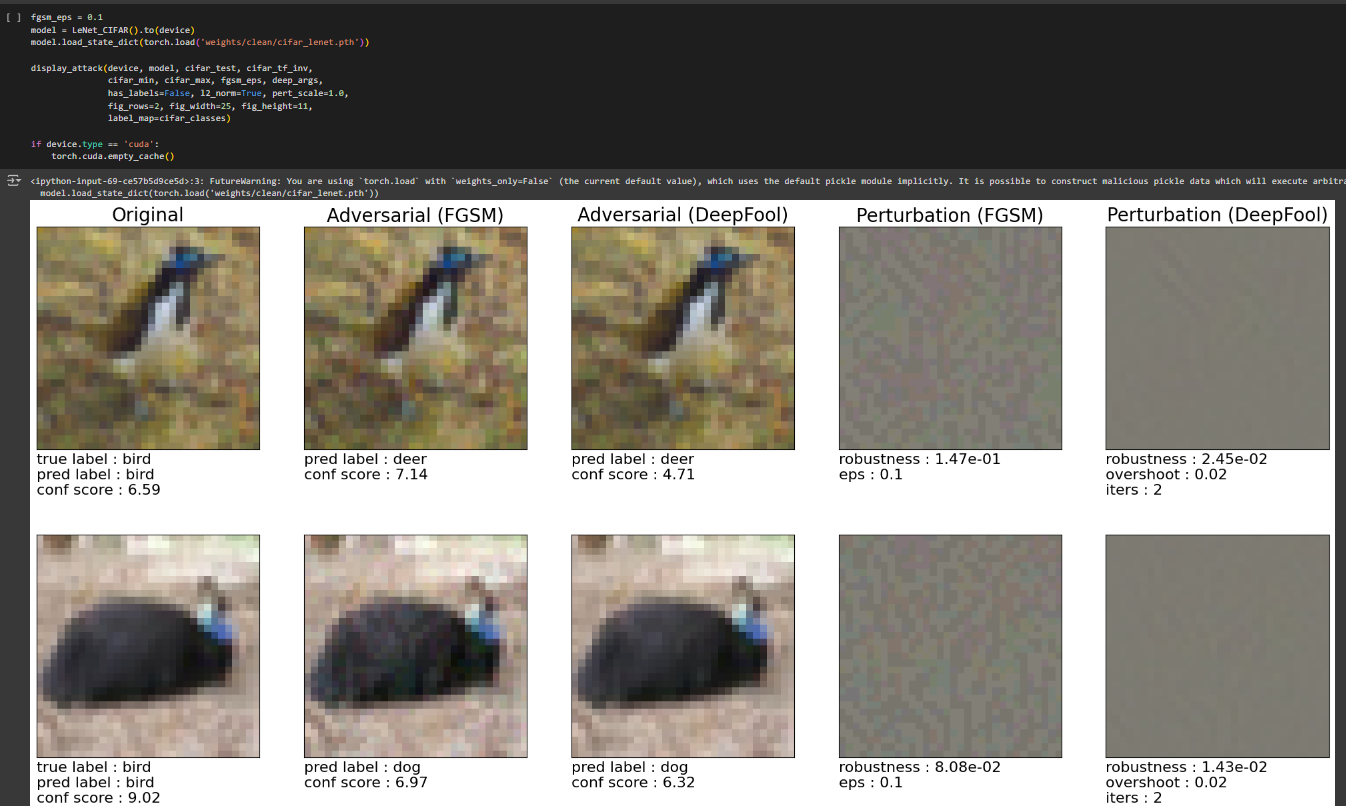


12. Выполнить оценку атакующих примеров для сетей:









**Результаты:**

Маленькие значения eps (например, 0.001, 0.02):

При маленьких значениях eps атака не сильно искажает данные. Градиенты, добавленные к исходным изображениям, настолько малы, что модель может легко отличить атакующие примеры от исходных.

Точность модели остается высокой. Например, при eps = 0.001 модель может практически не терять точности. Это значит, что такая атака не является эффективной, так как она минимально искажает изображения.

Низкие значения eps создают почти незаметные искажения, и модель сохраняет свою устойчивость.

Средние значения eps (например, 0.02, 0.5):

Как только значение eps увеличивается, модель начинает сталкиваться с трудностями. Атака становится более заметной, и ошибки начинают накапливаться. При eps = 0.02, некоторые примеры начинают давать ложные предсказания, хотя точность все еще остается достаточно высокой.

Например, при eps = 0.02, модель может потерять несколько процентов точности. При eps = 0.5 точность может значительно упасть.

Средние значения eps демонстрируют, как легко можно запутать модель с увеличением силы атаки.

Большие значения eps (например, 0.9, 10):

При очень больших значениях eps искажения становятся настолько сильными, что модель практически не может отличить атакующие примеры от случайного шума. Изображения сильно деформированы и больше не представляют собой исходные данные.

При eps = 0.9 точность модели может упасть до критического уровня.

При экстремально больших значениях, таких как eps = 10, модель перестает предсказывать правильно, и точность падает практически до нуля.

Сильно большие значения eps приводят к разрушительным атакам, которые полностью разрушают представление модели об исходных данных. Это показывает, насколько уязвимы модели при больших искажениях данных.

**Заключение**

В результате выполнения лабораторной работы было выявлено, что маленькие значения fgsm\_eps сохраняют стойкость сетей к атакам, и ошибки классификации остаются низкими. При увеличении fgsm\_eps сети становятся более уязвимыми к атакам и допускают больше ошибок классификации. Для сети FC LeNet на датасете MNIST и для сети NiN LeNEt на датасете CIFAR не наблюдается отсутствие влияния параметра fgsm\_eps. Наоборот, параметр fgsm\_eps оказывает существенное влияние на стойкость сетей к атакам.

Ссылка на colab: https://colab.research.google.com/drive/1qJKmab4-YISIF08hKEzgQ\_MYY4X-6gFZ?usp=sharing