|  |  |  |
| --- | --- | --- |
|  |  |  |
| 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»**

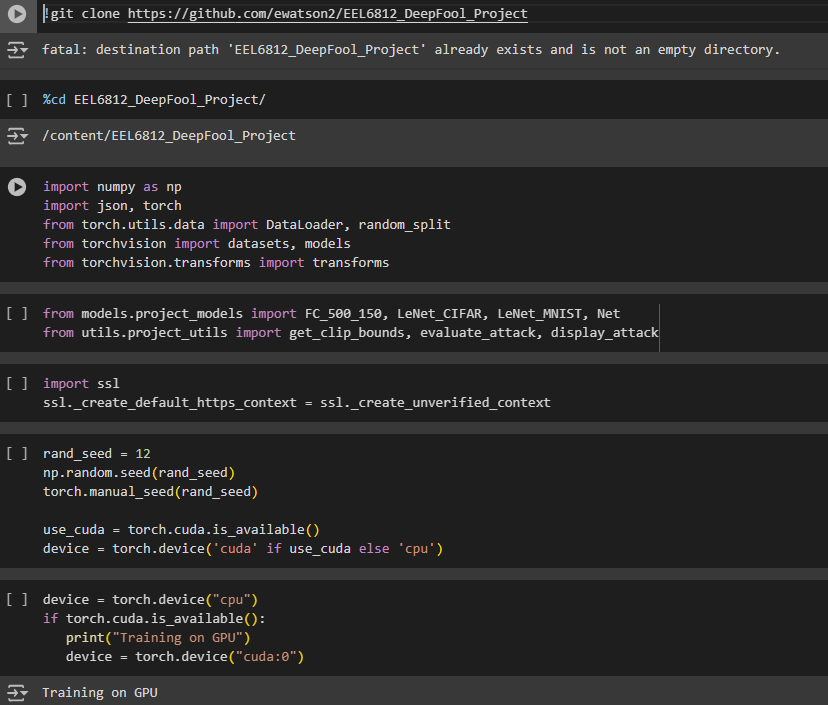
Выполнил(а)

Студент 1 курса

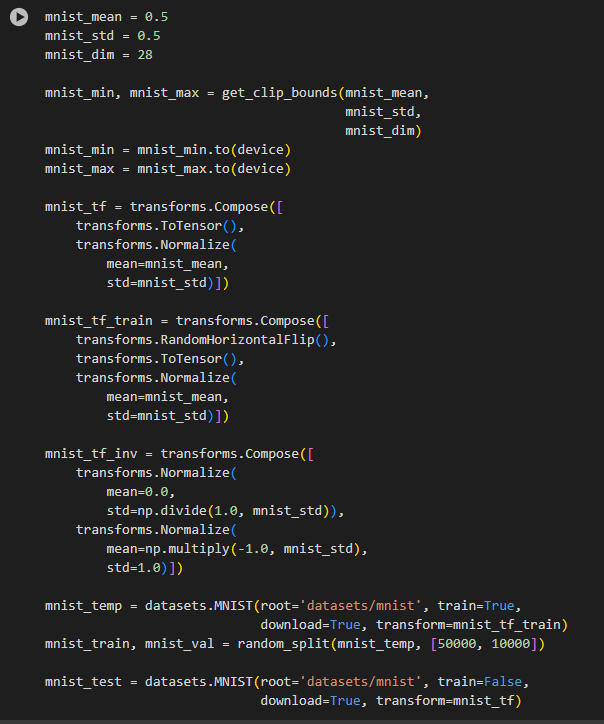
Группы ББМО-02-23

Колонин Артём Николаевич

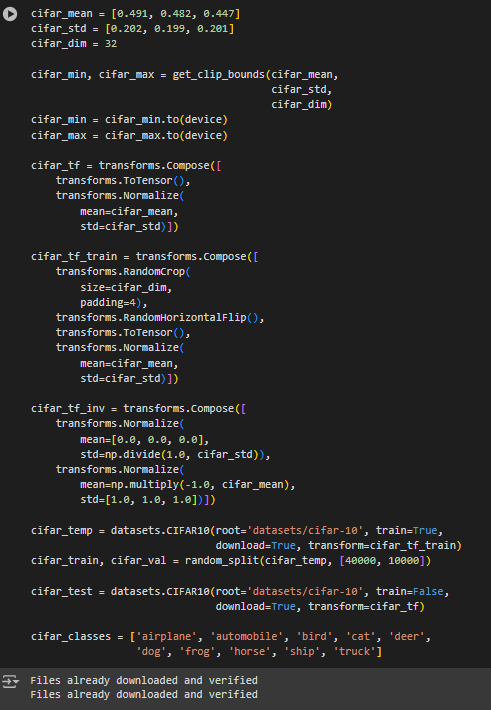
Скачиваем проект, переключаемся на GPU для ускорения вычисений.



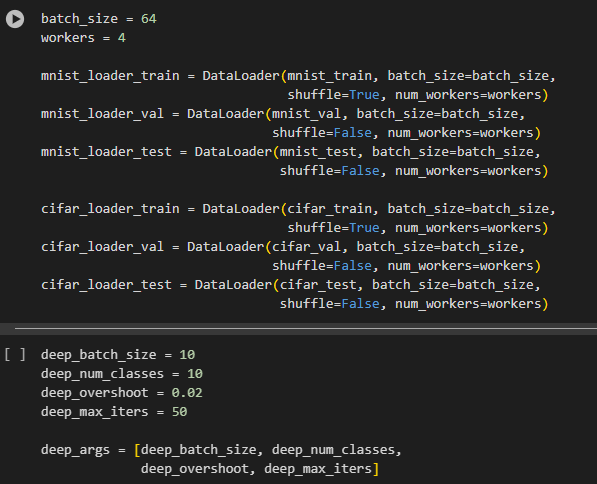
Загружаем датасет MNIST



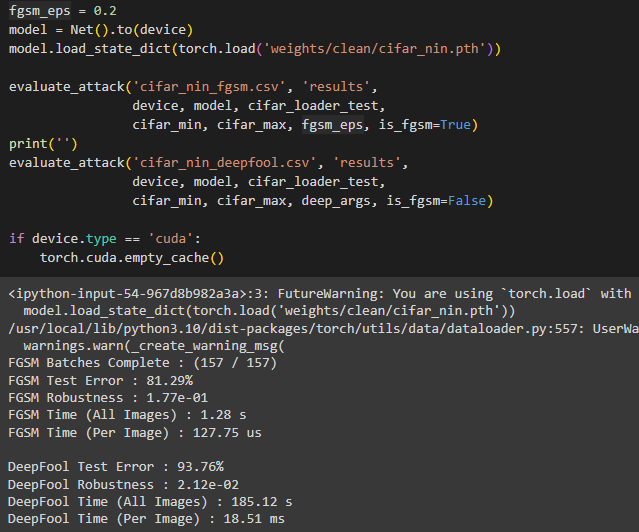
Загружаем датасет CIFAR



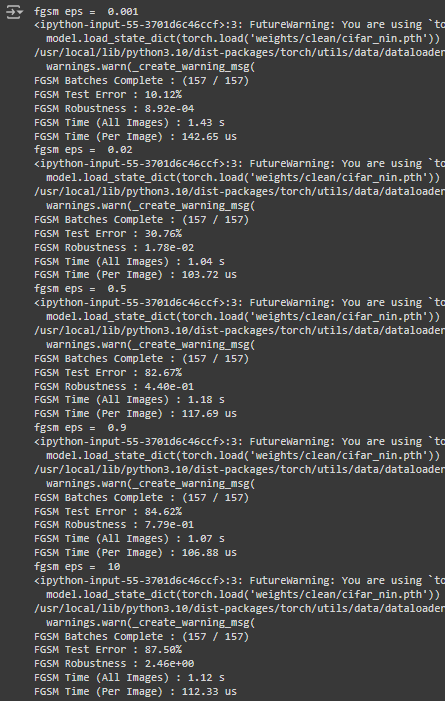
Разбиваем датасеты на выборки, задаём гиперпараметры для Deepfool



Нейросеть Network in network на датасете Cifar, атака Deepfool привела к большей ошибке, чем атака FGSM



Рассмотрим различные значения переменной eps при атаке FGSM и сравним ошибки нейросети



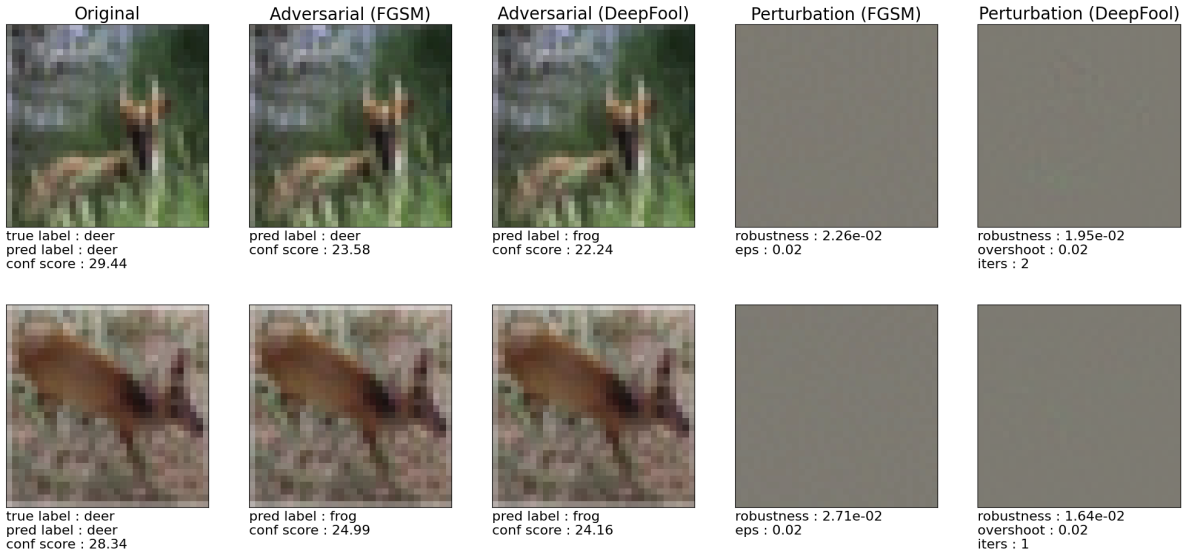
При значении eps <= 0.02 ошибка была довольно мала, 30%, а на 0.5 уже 82%. Оптимальное значение находится где-то между этими примерами.

Рассмотрим, как значение переменной eps влияет на входные изображения

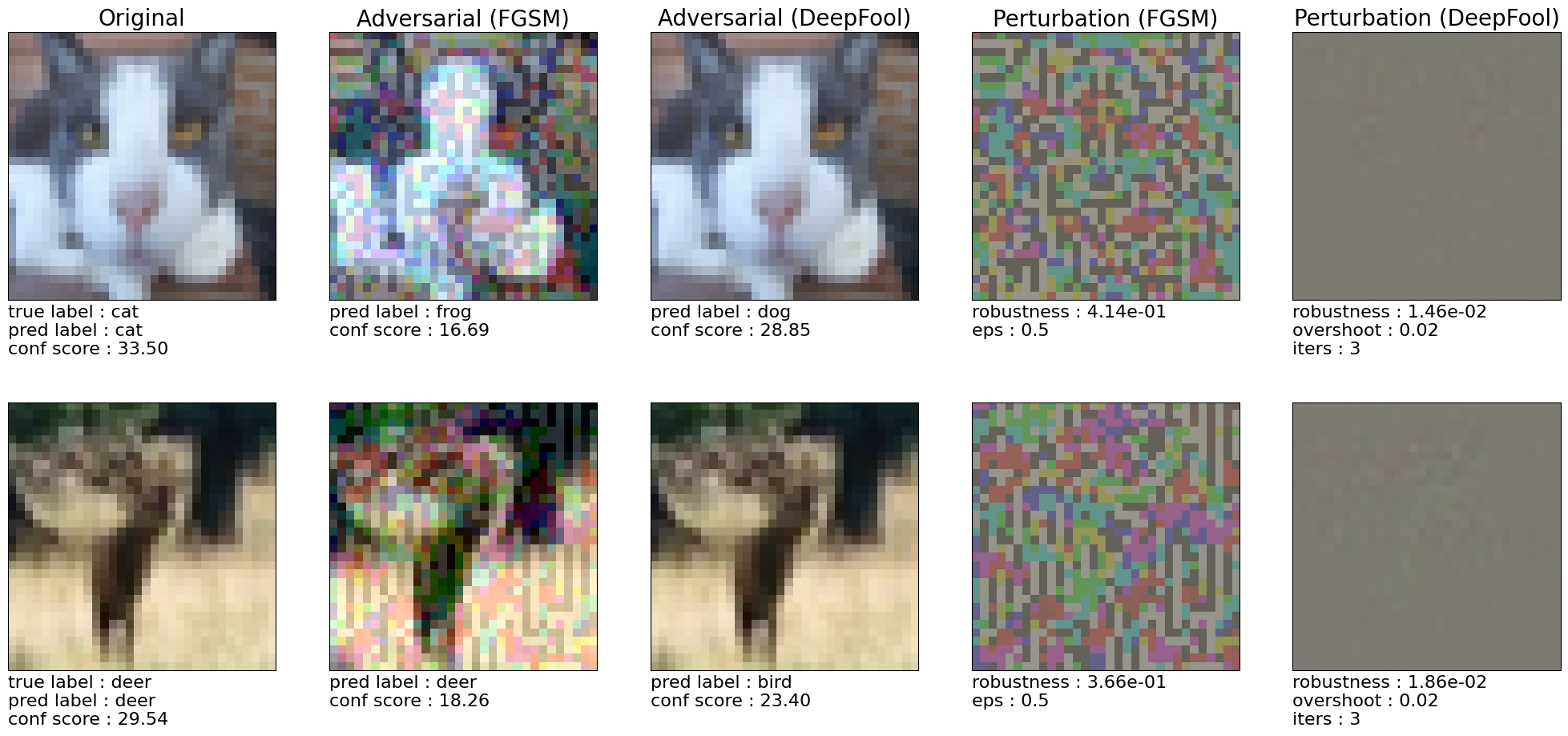
0.001



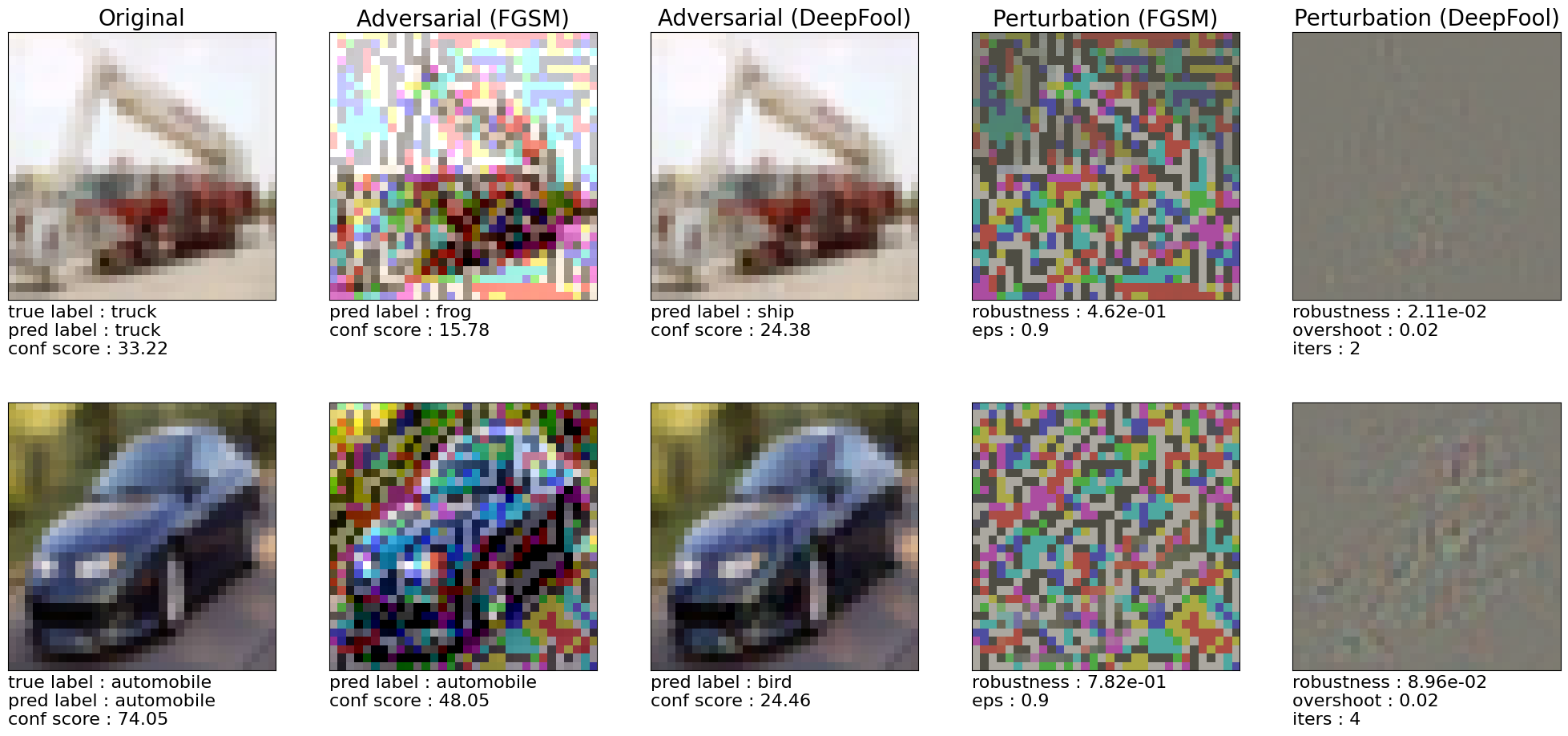
0.02



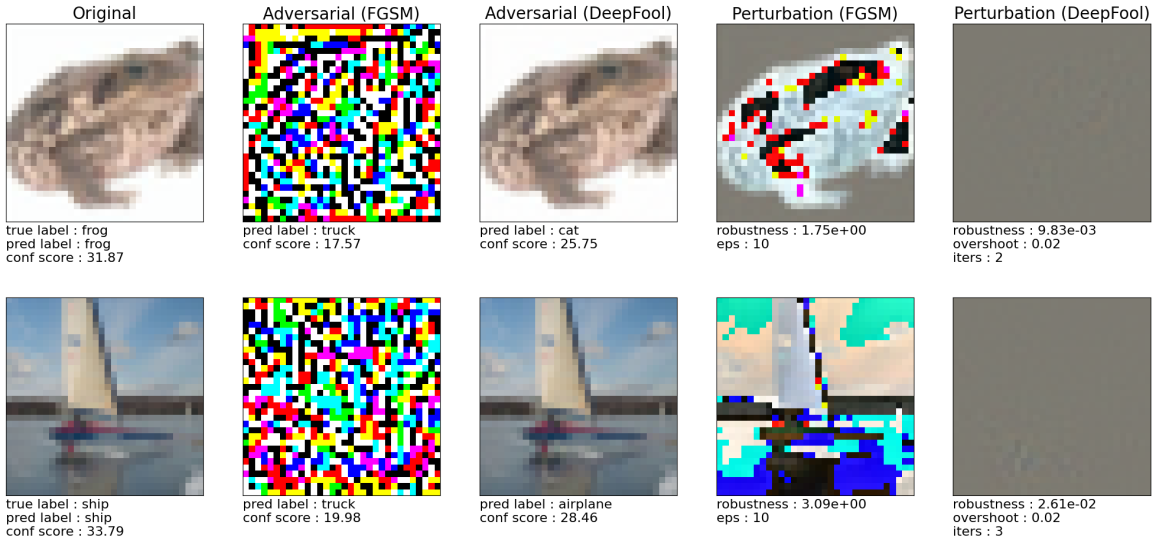
0.5



0.9

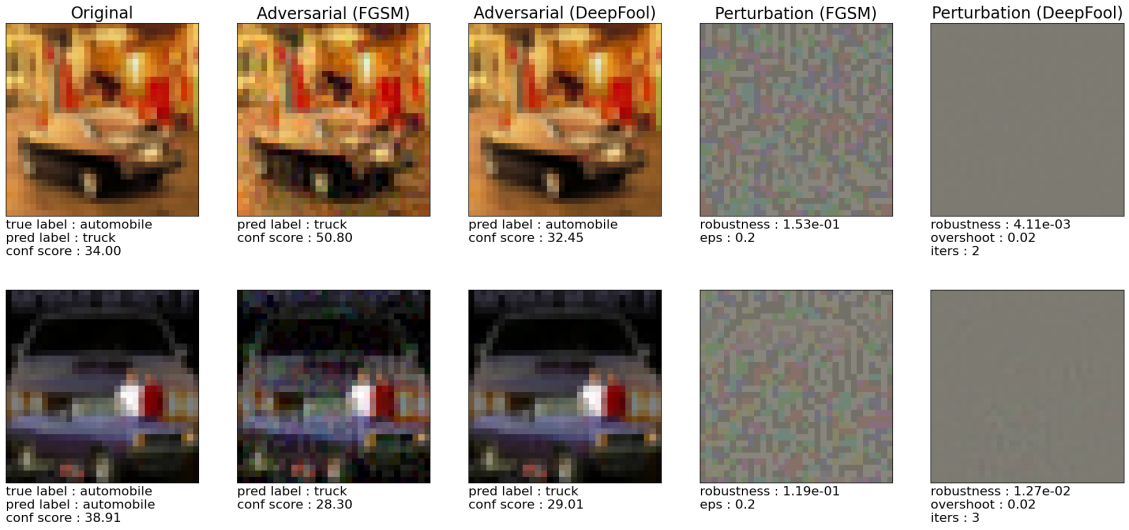


10



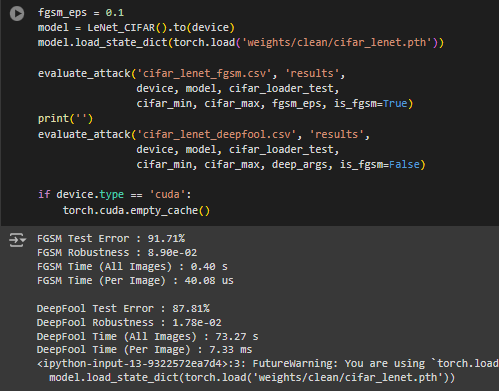
Значения меньше 0.1 почти не несут никаких искажений и не сильно ухудшают работу нейросети. Однако 0.5 уже заметно изменяет изображение. Возьмём значение 0.2.

0.2

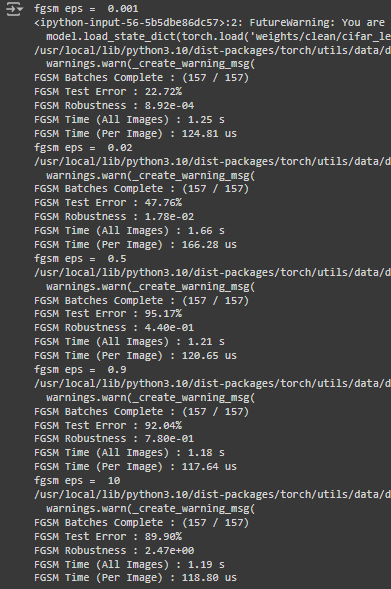


Искажения почти не заметны, а ошибка составляет 81%.

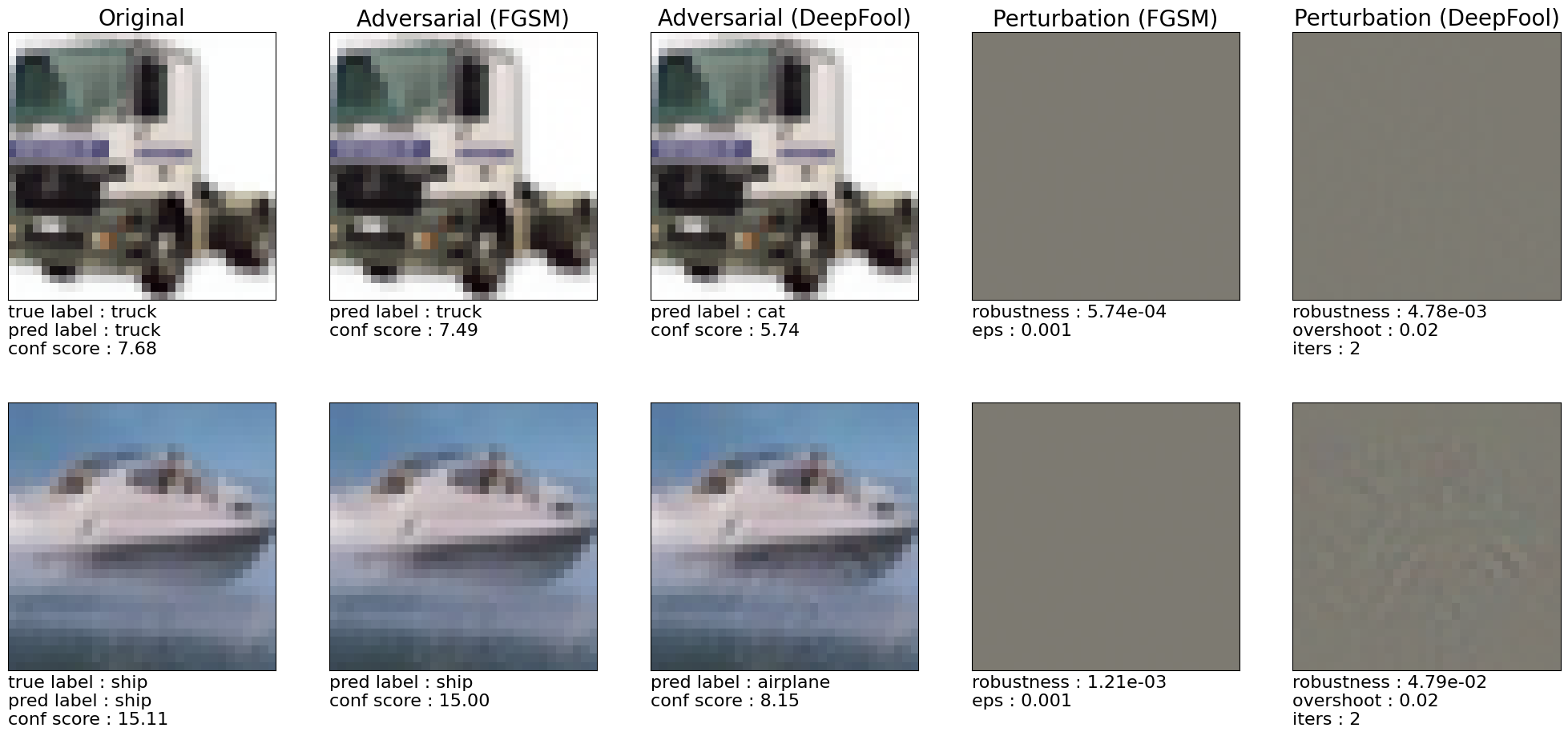
Теперь рассмотрим нейросеть Lenet на том же датасете cifar



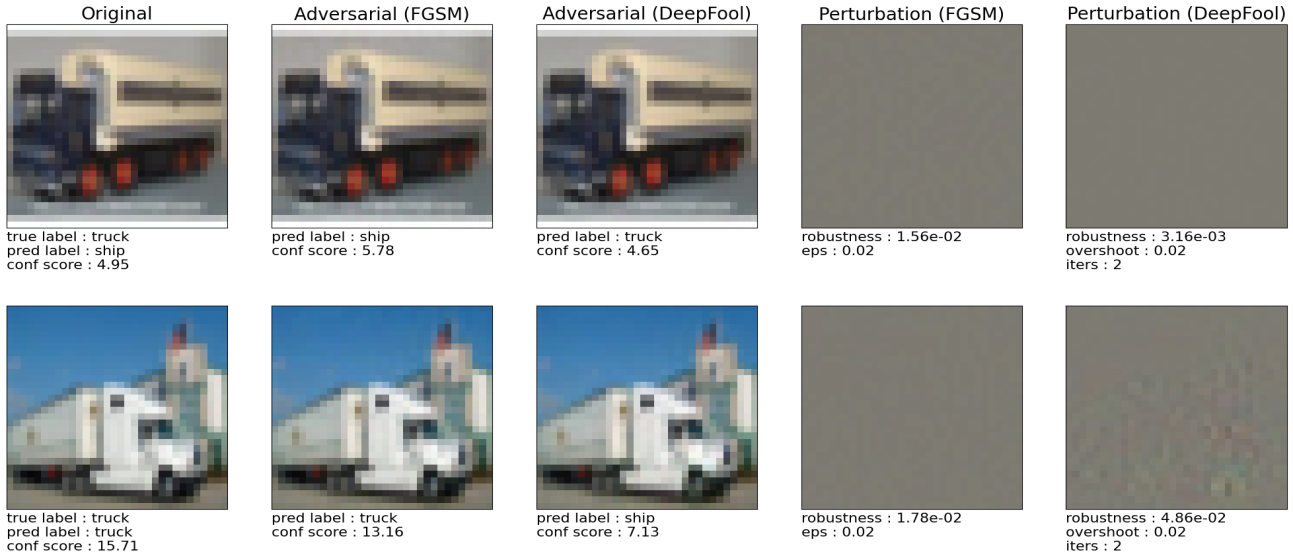
Рассмотрим такие же значения переменной eps и их влияния на точность и искажения.



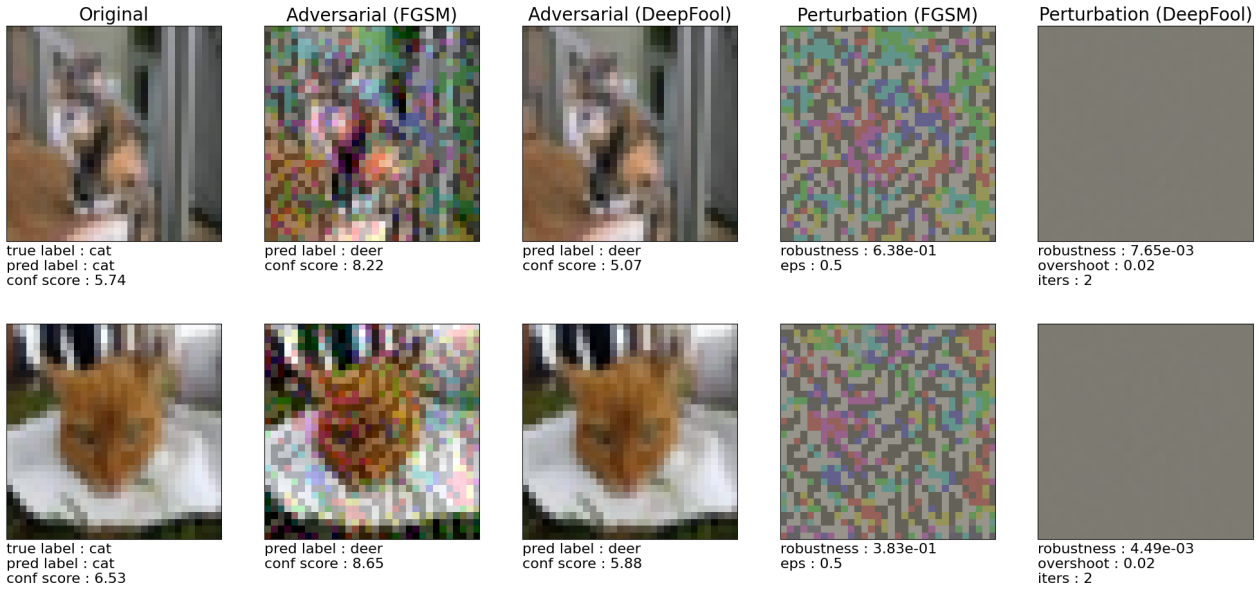
0.001



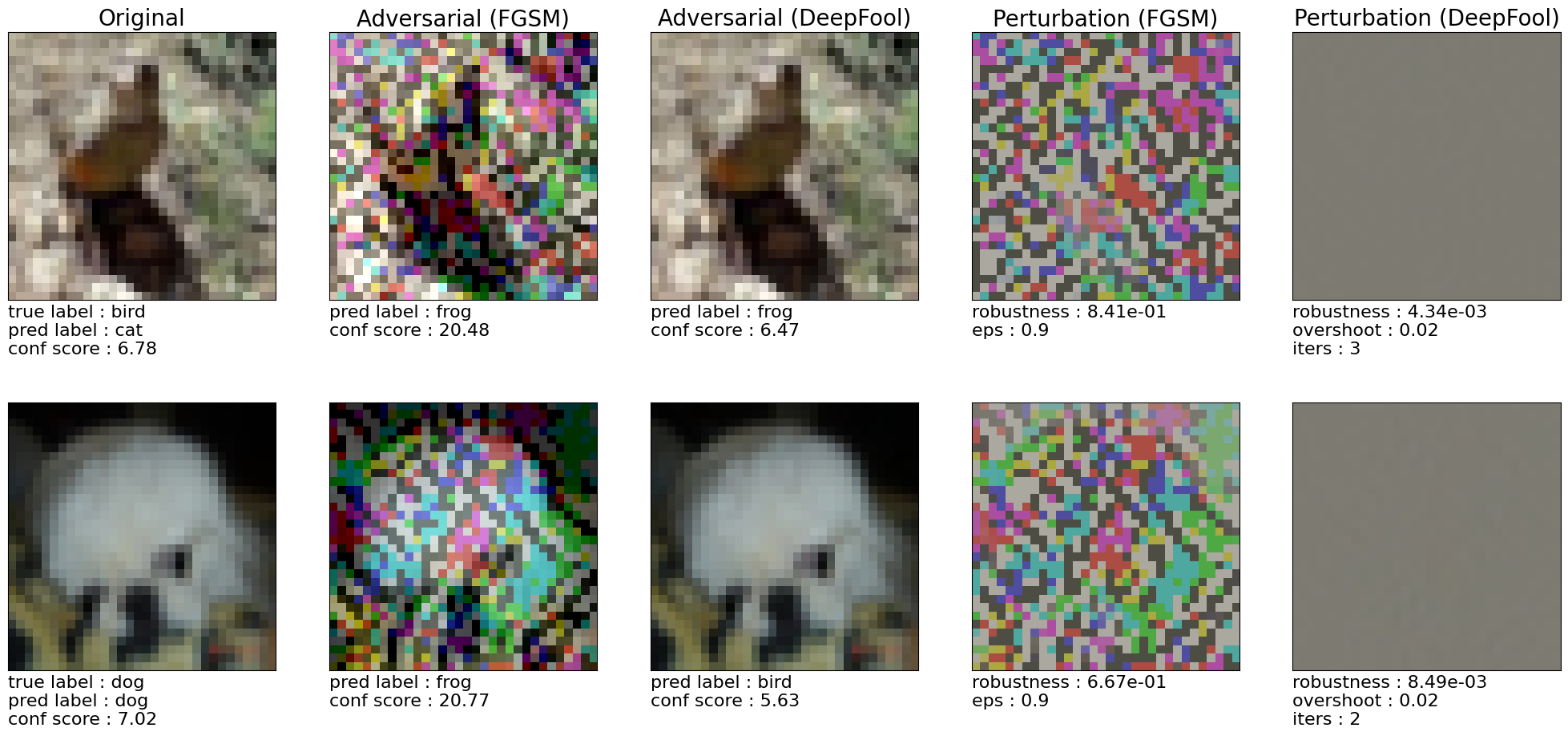
0.02



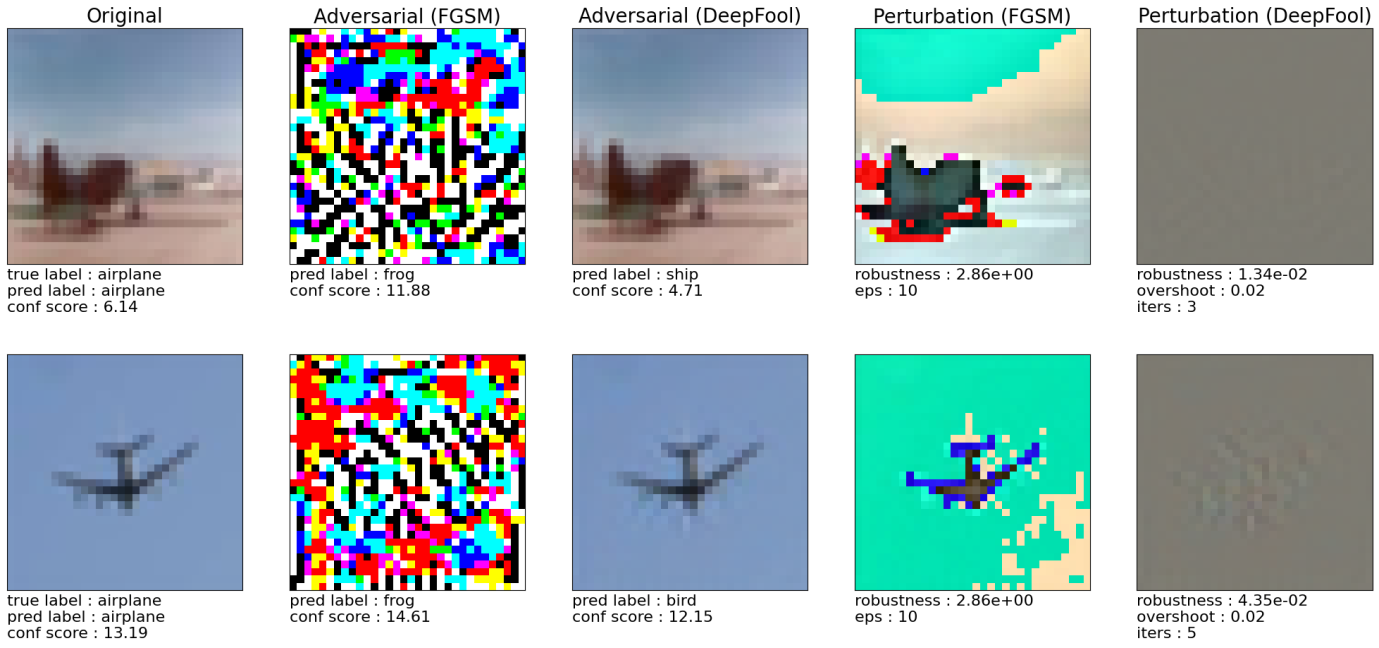
0.5



0.9

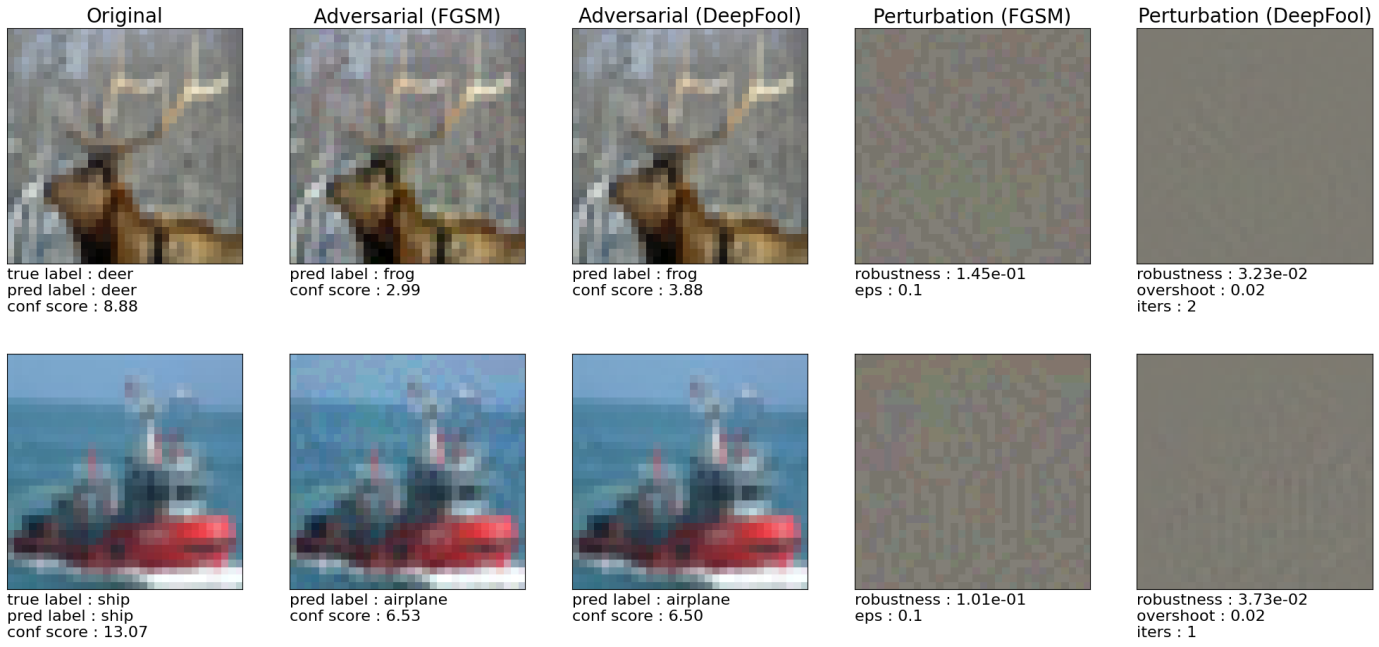


10

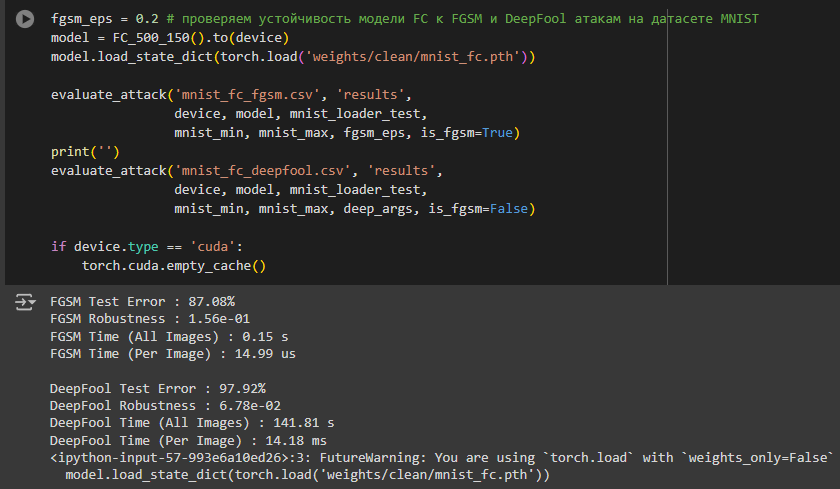


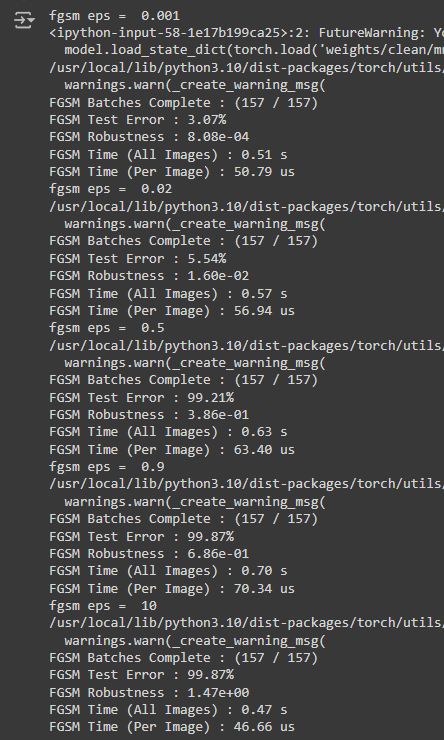
Ситуация аналогична предыдущей нейросети, но при eps > 0.5 ошибка наоборот начинает уменьшаться. На этот раз возьмём eps=0.1 и также получим несильные искажения при ошибке в 91%

0.1

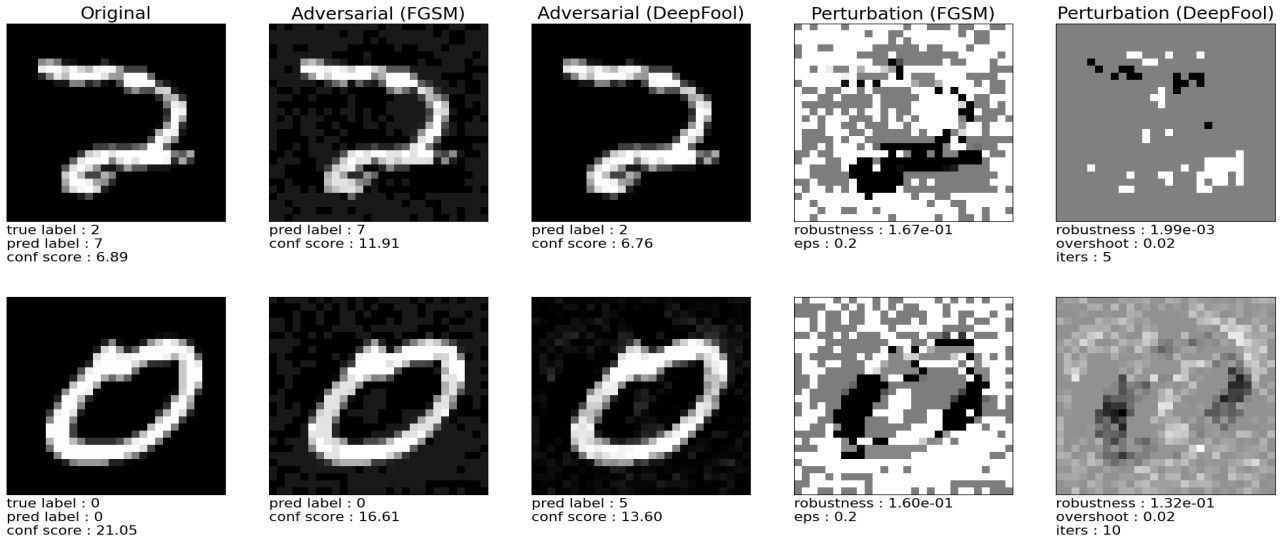


Теперь рассмотрим нейросеть FC на датасете MNIST, проведём точно такие же исследования

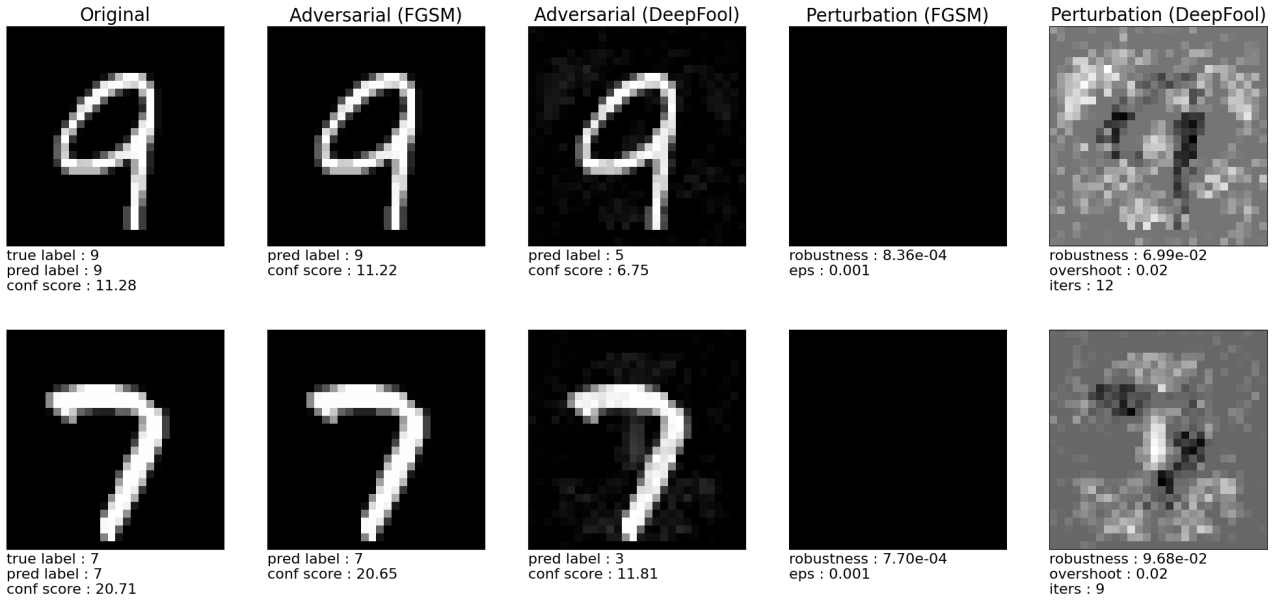




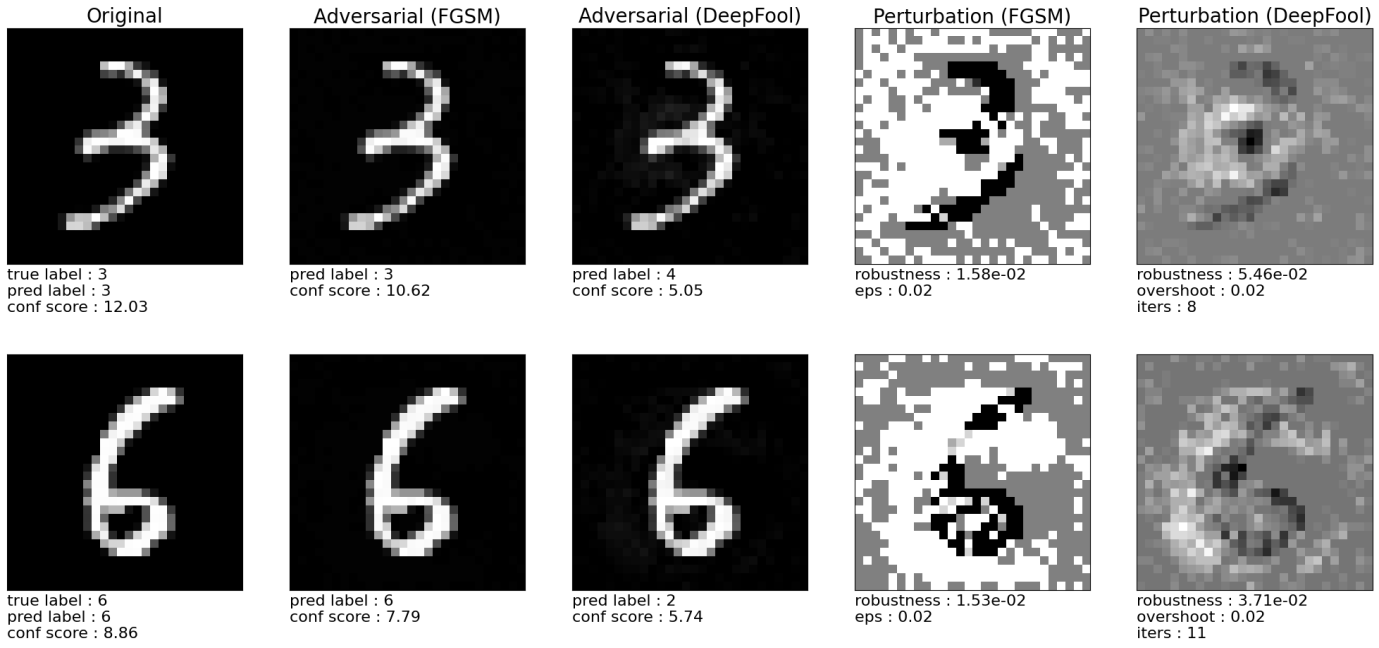
0.2



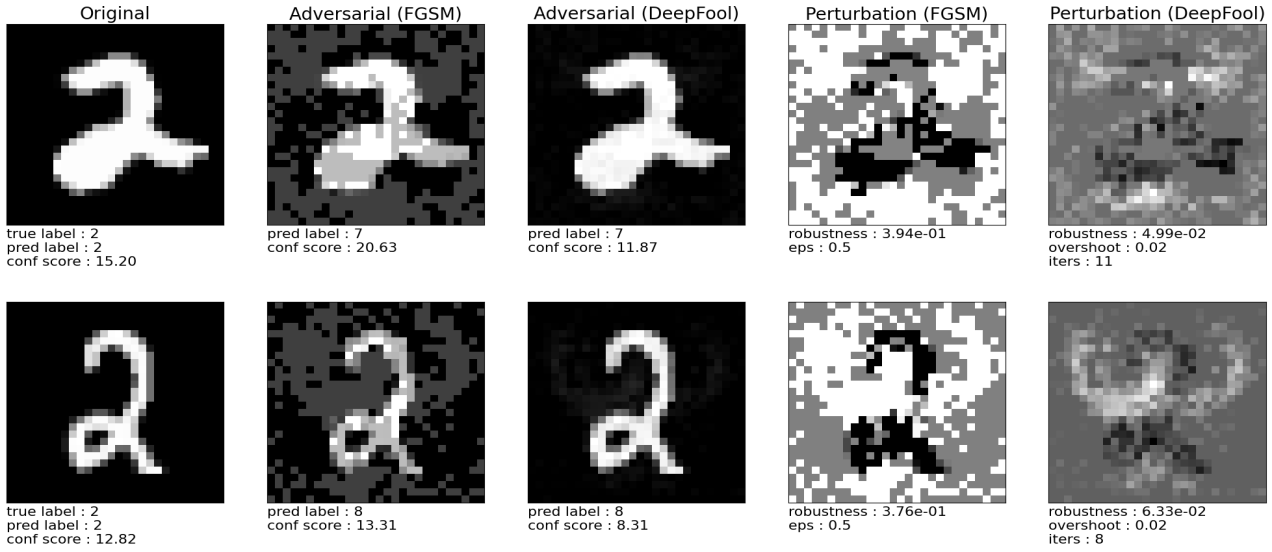
0.001



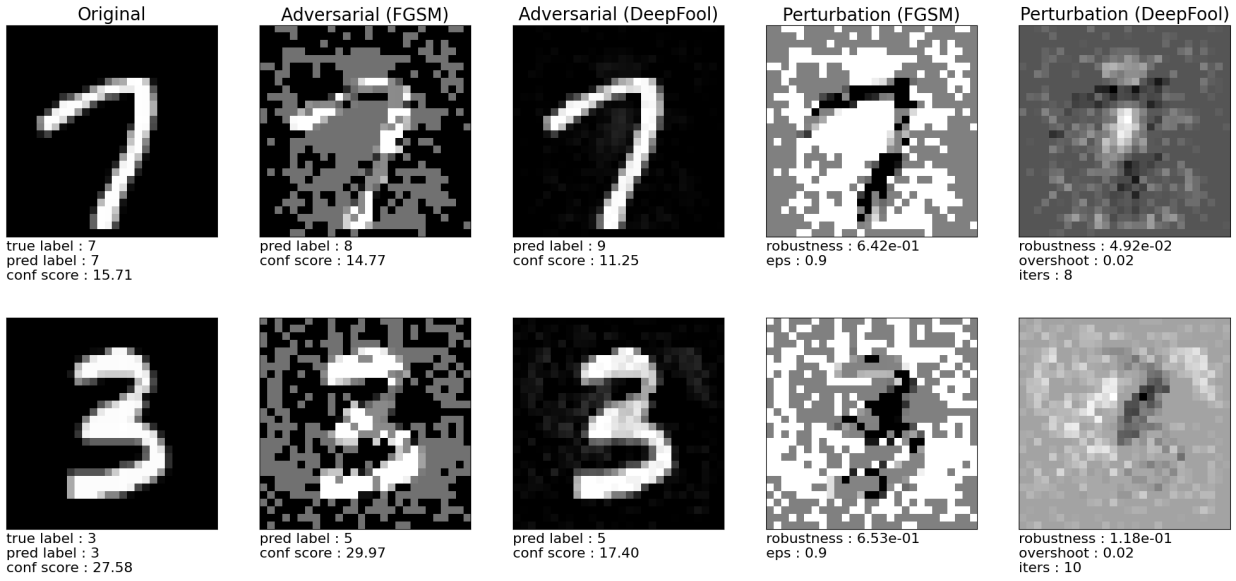
0.02



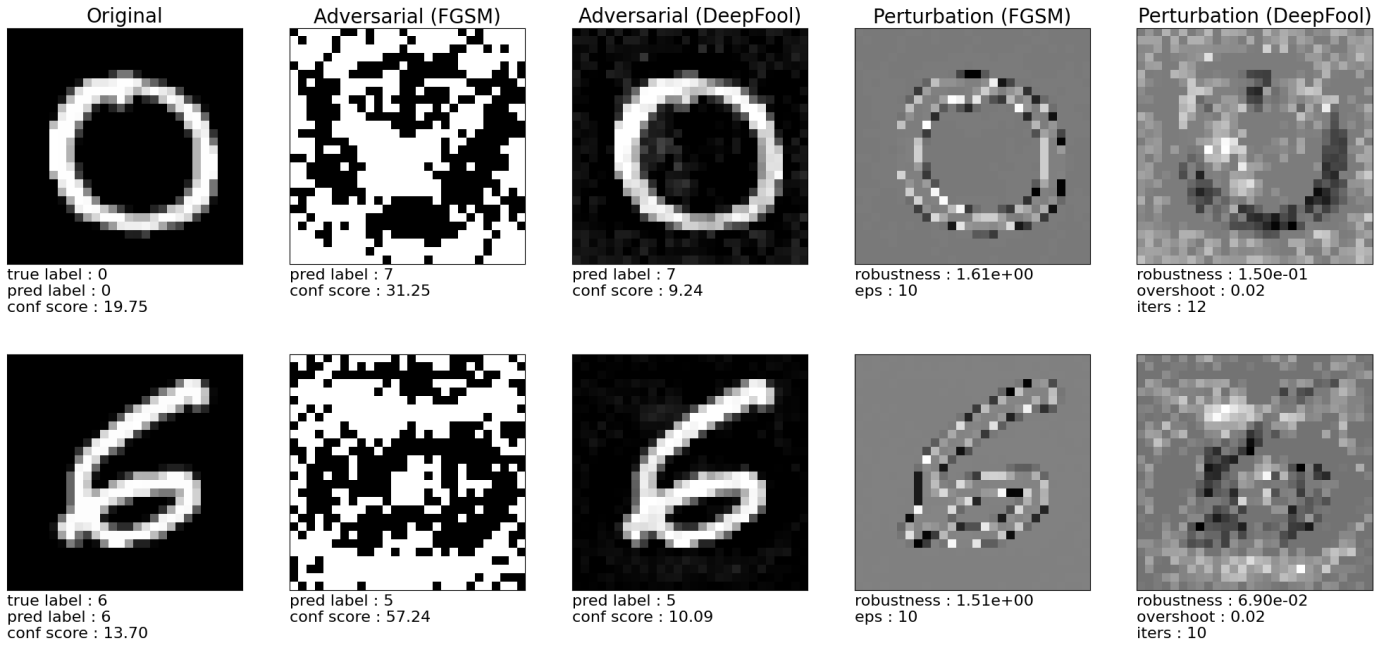
0.5



0.9

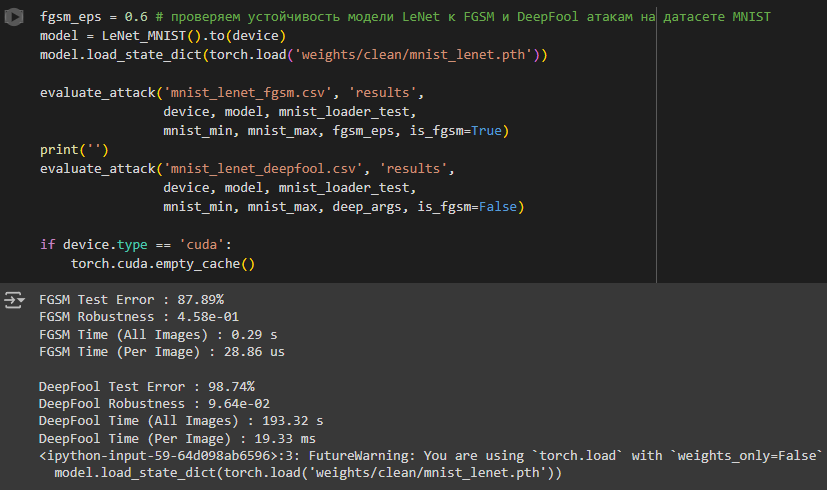


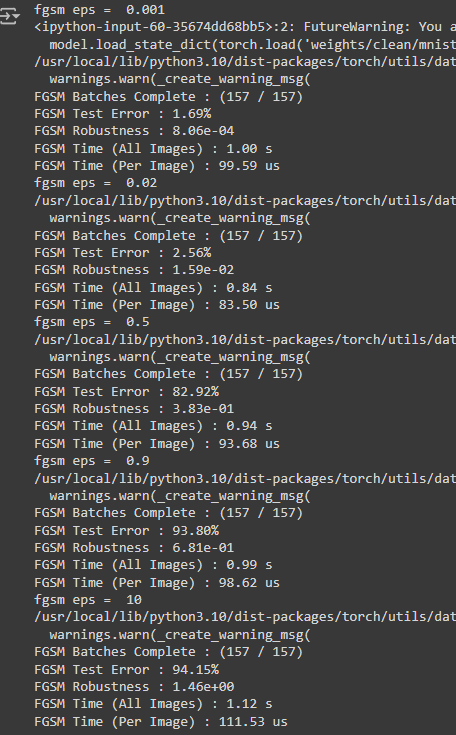
10



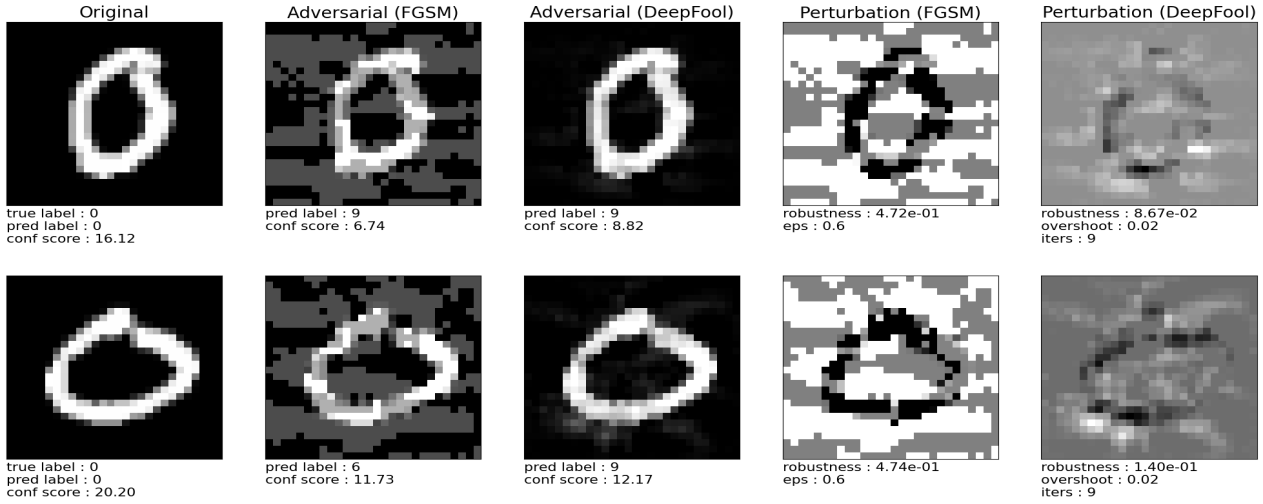
Снова слишком большие значения eps приводят к сильно искажённым изображениям. При eps=0.2 искажения едва заметны, а точность падает до 87%.

Последняя нейросеть - Lenet на датасете MNIST

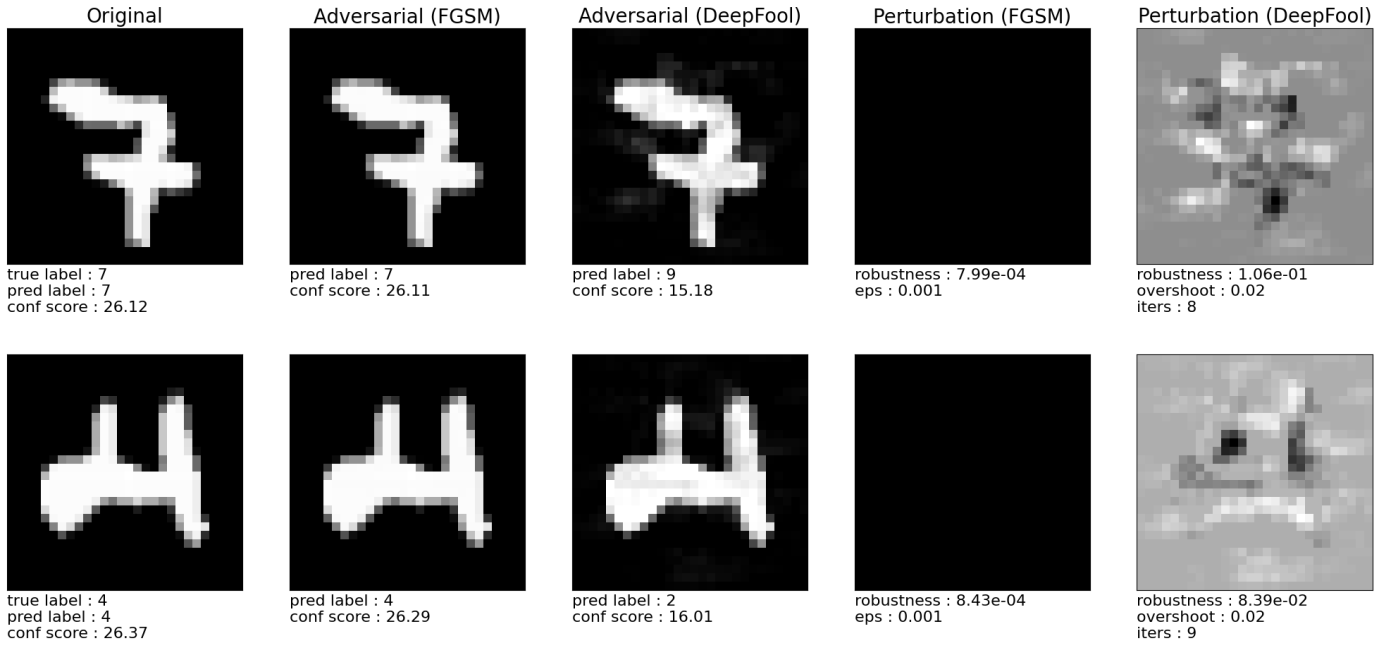




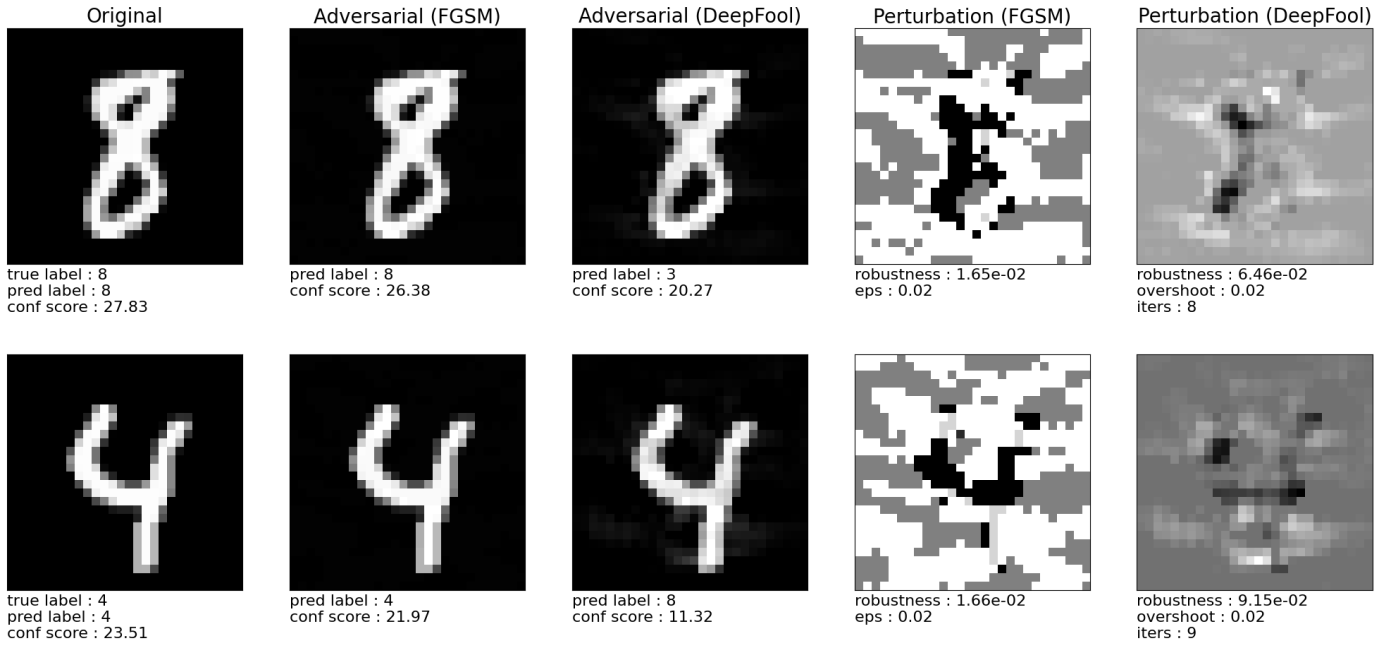
0.6



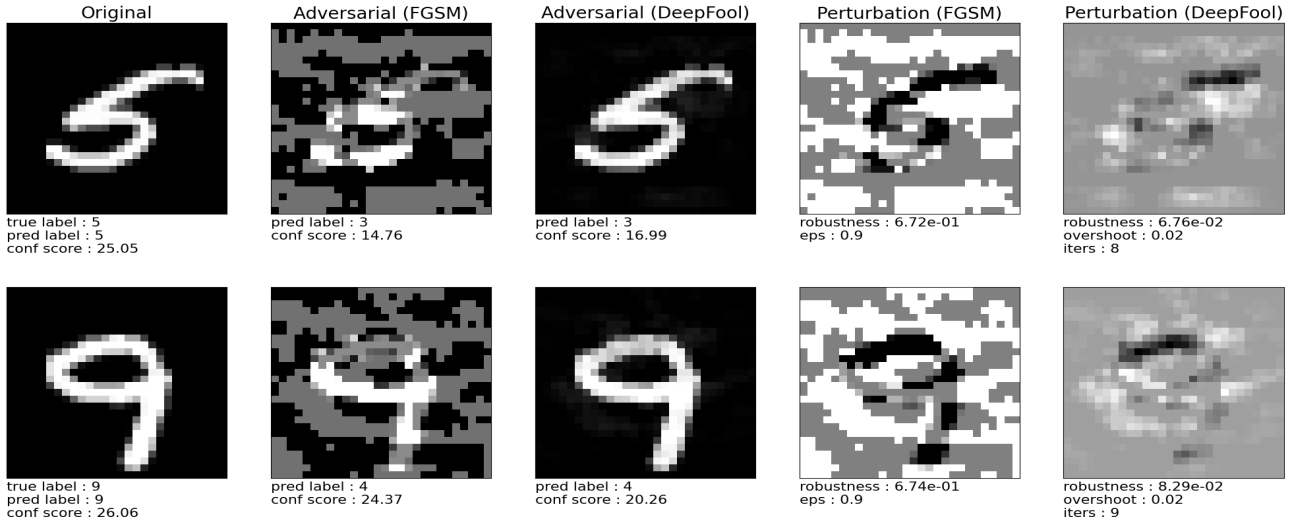
0.001



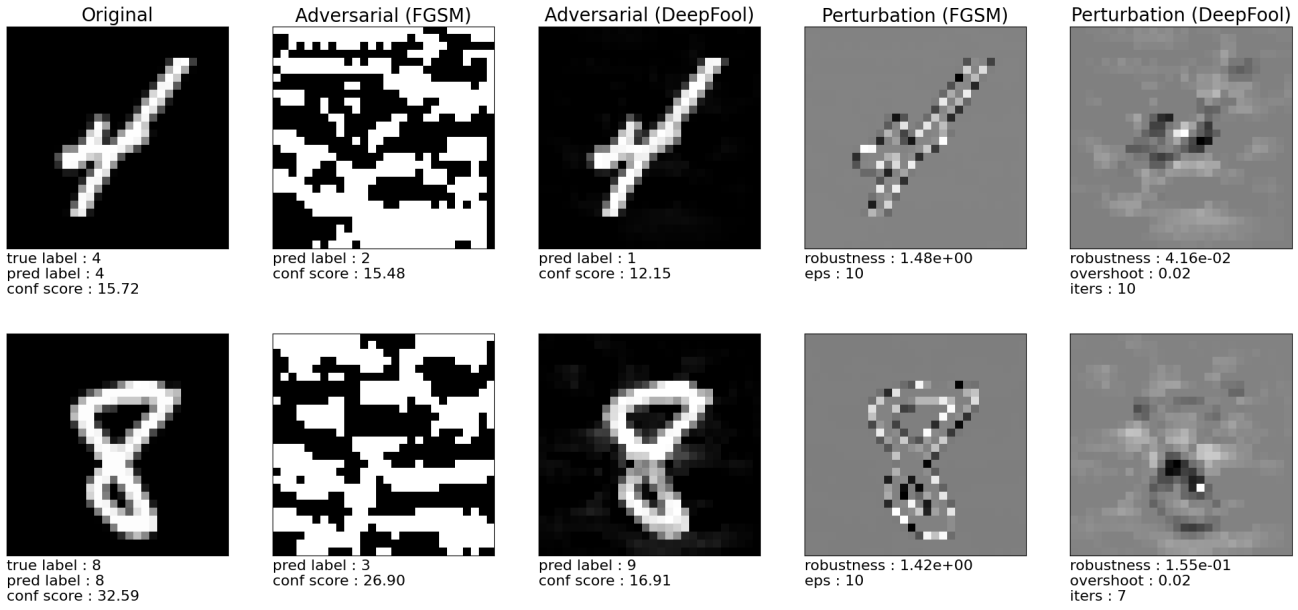
0.02



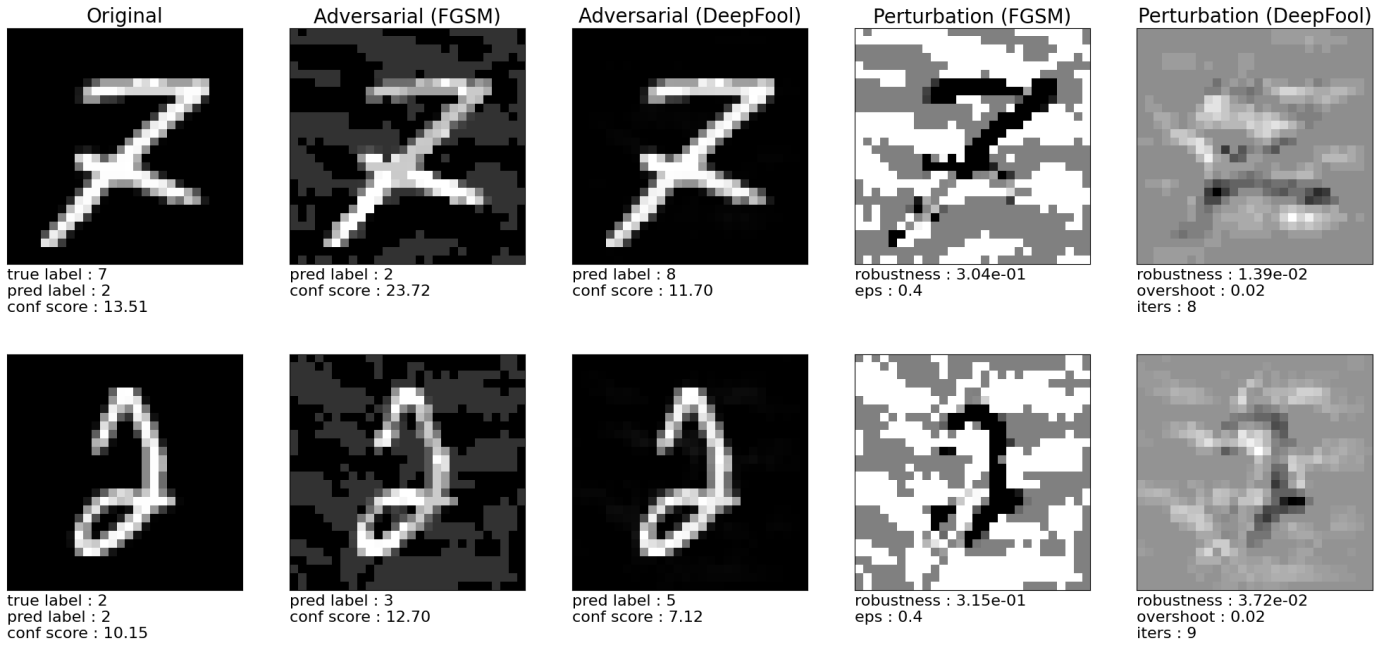
0.9



10



0.4



При eps=0.6 ошибка составляет 87%, хотя искажения весьма заметны.

Таким образом при увеличении eps увеличиваются искажения на исходных изображениях и снижается точность нейронной сети. Оптимальное значение не должно быть слишком большим, иначе атака будет слишком заметна.

Самая уязвимая для атак нейросеть из рассмотренных – FC на датасете MNIST, так как у этой нейросети больше остальных росла ошибка при атаке. Однако, остальные нейросети также достаточно уязвимы – при правильном подборе параметров их точность можно уменьшить на 80% и больше.