|  |
| --- |
| https://lh6.googleusercontent.com/4WfKLPaR-DoCa9xcfJIuSyzkqJT0R_6Hdu7tpdJ6XSo2x1S7gKz5Z72FtFU_YHTcugtZeiO0GGQZDF0HevpOm7L2RlF5WAVzEDZ71HK35mDFG6eAxSSjQVC88ArmpkmEXjU4rKBM  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»

Отчет по лабораторной работе 3

по дисциплине

««Анализ защищенности систем искусственного интеллекта»»

**Выполнил**

                                        Студент 2 курса: Стельмах Н.Е.

Группы: ББМО-02-22

Проверил:

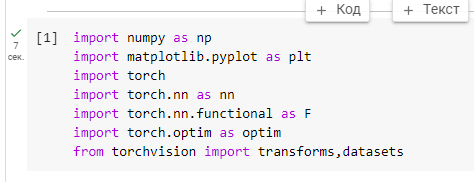
Спирин А.А

Москва 2022 г.

**Изучение методов защиты от атак на модели НС**

**Защитная дистилляция**

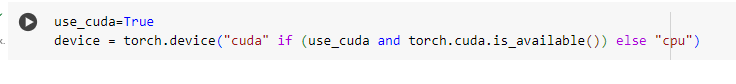
1. Выполнить импорт необходимых библиотек.



2. Задать нормализующие преобразования? загрузить набор данных (MNIST), разбить данные на подвыборки

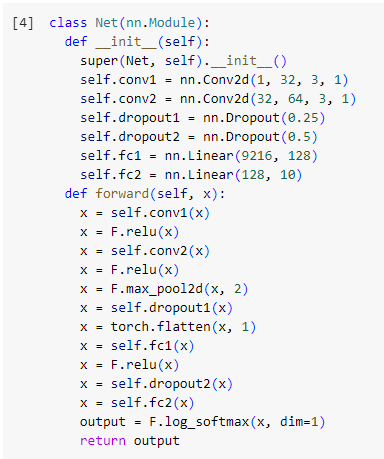


3. Настроить использование графического ускорителя (если возможно)

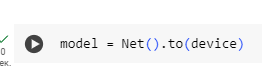


**Создание атак на модель НС**

4. Создать класс НС на основе фреймворка torch



5. Проверить работоспособность созданного класса НС



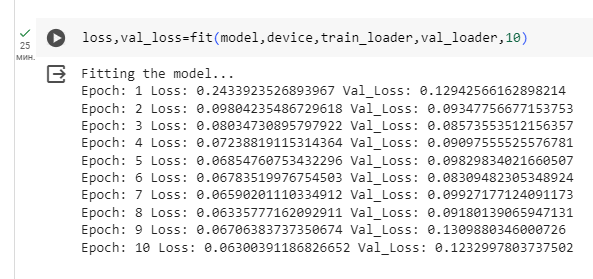
6. Создать оптимизатор, функцию потерь и трейнер сети.



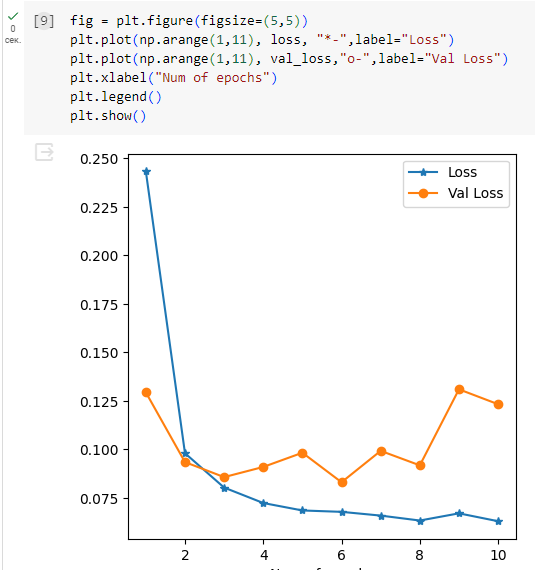
7. Определить функцию обучения сети



8. Обучить модель



9. Построить графики потерь при обучении и валидации в зависимости от эпохи



10. Создать функции атак FGSM, I-FGSM, MI-FGSM

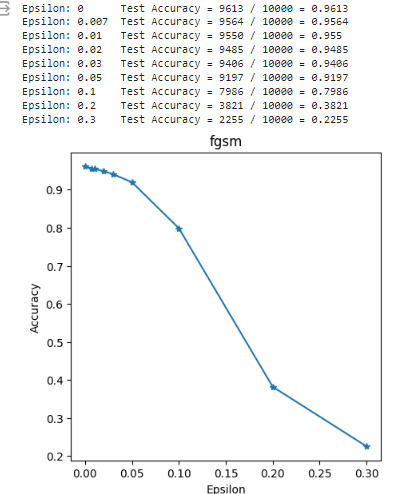


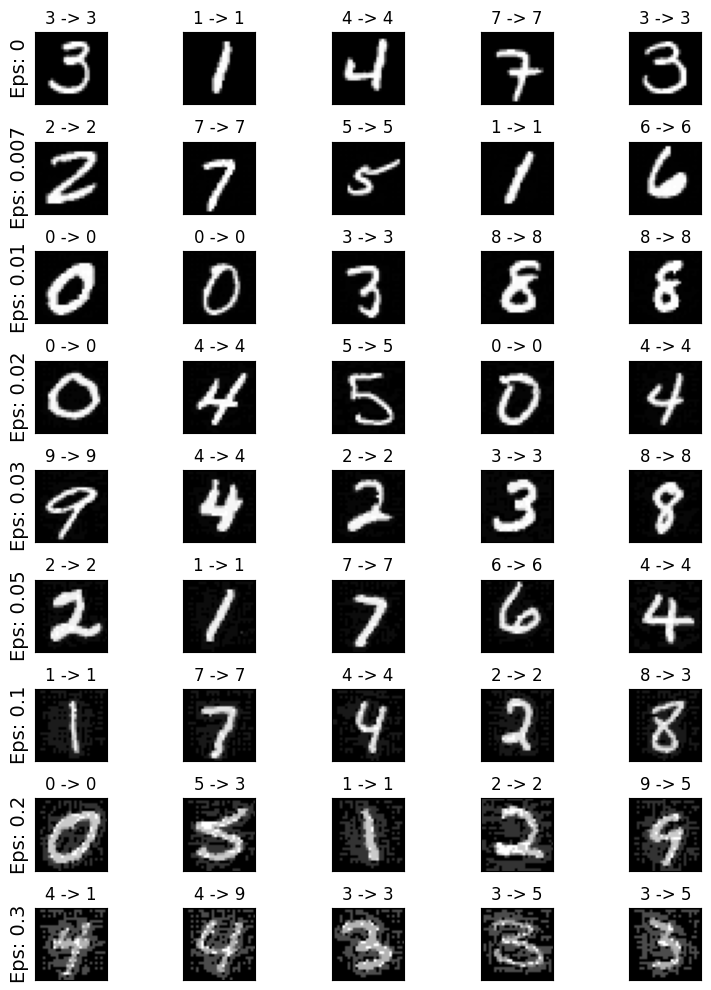
11. Создать функцию проверки

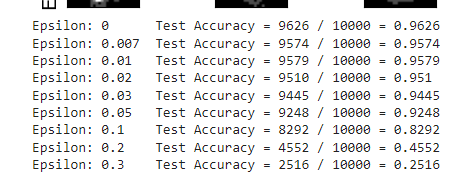


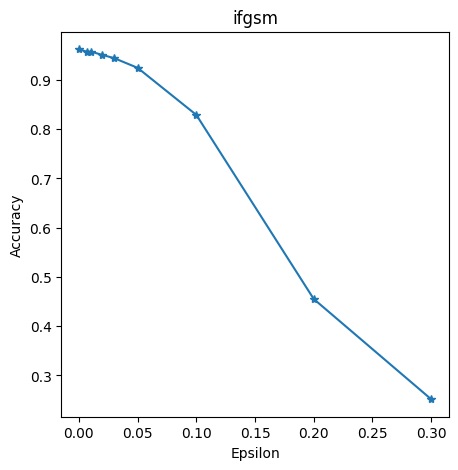
12. Построить графики успешности атак(Accuracy/эпсилон) и примеры выполненных атак в зависимости от степени возмущения epsilon:

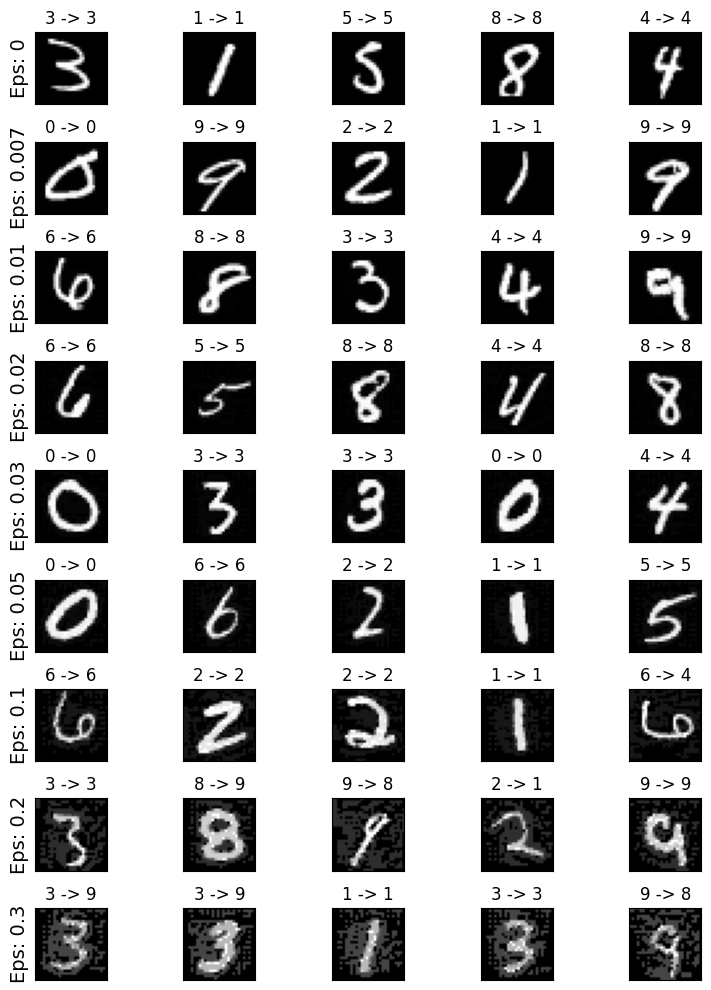


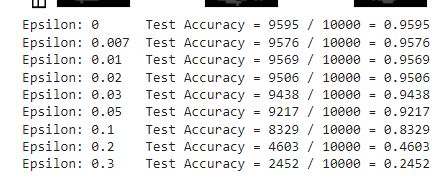


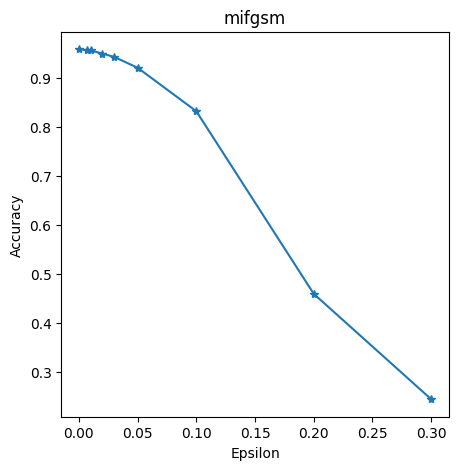


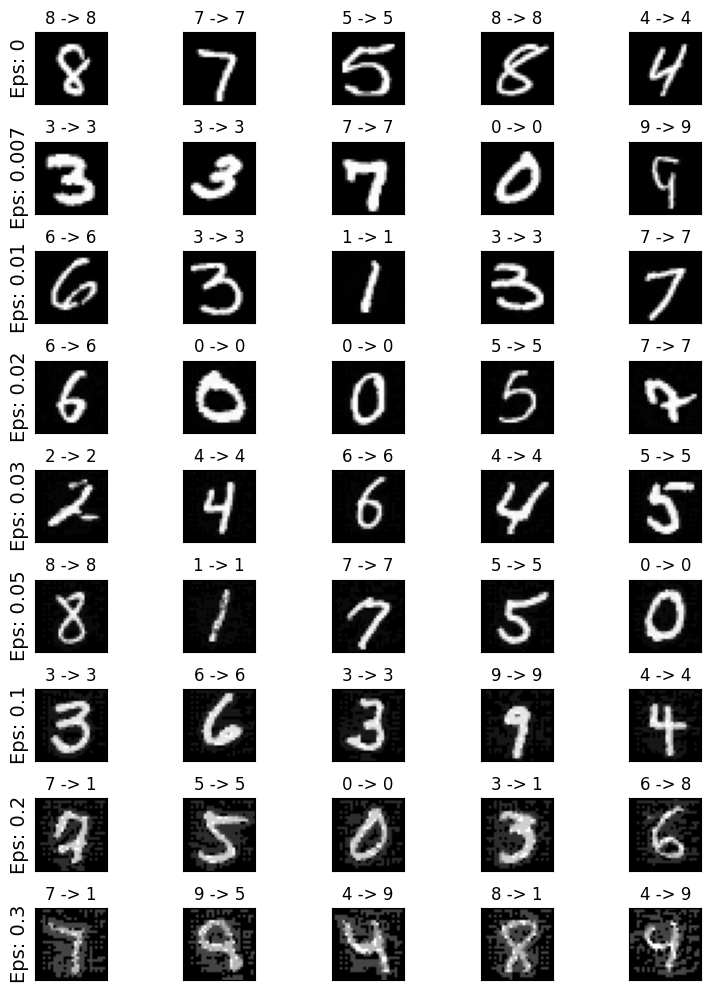






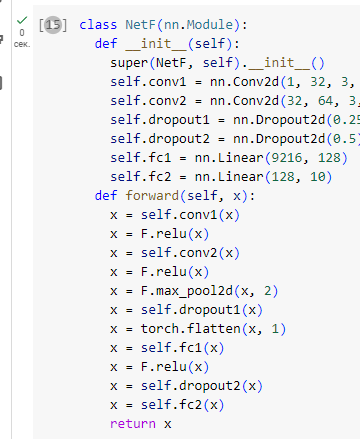






**Защита от атак**

13. Создать 2 класса НС





14. Переопределить функцию обучения и тестирования



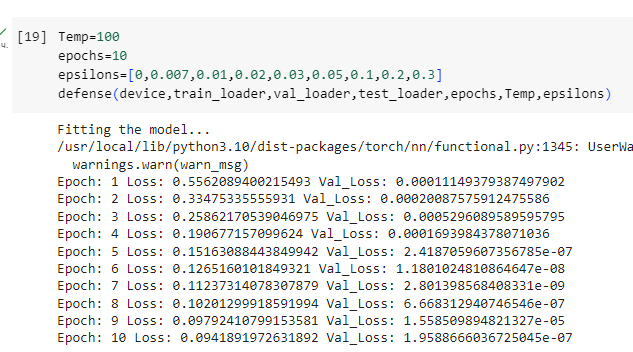


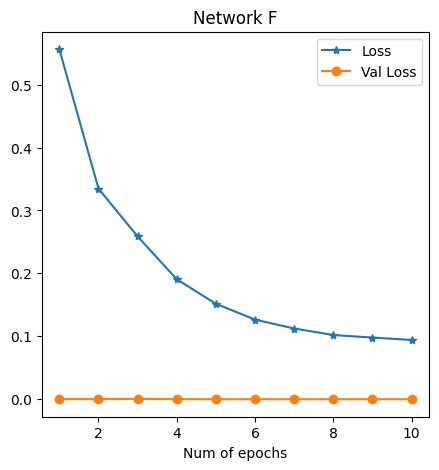
15. Создать функцию защиты методом дистилляции

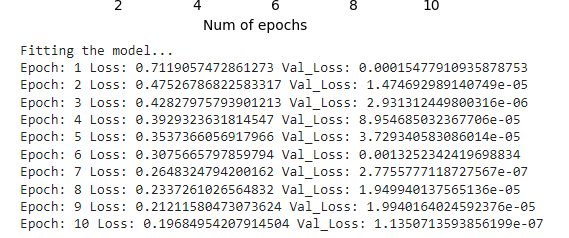


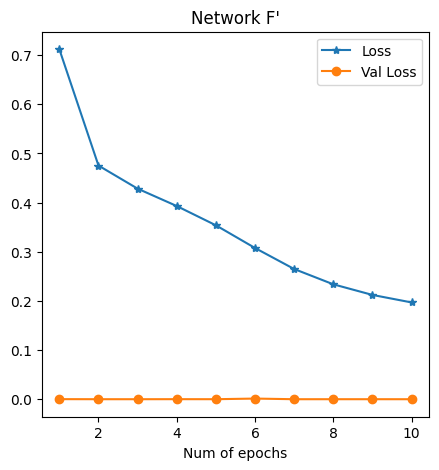


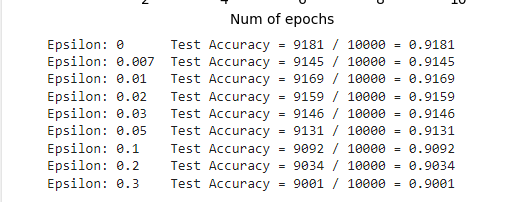
16. Получить результаты оценки защищенных сетей

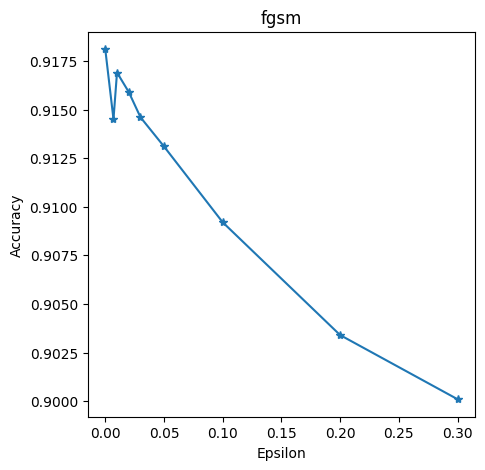


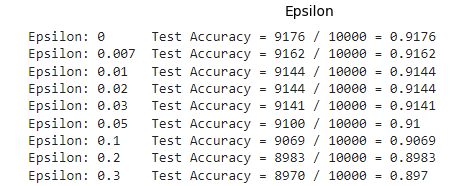


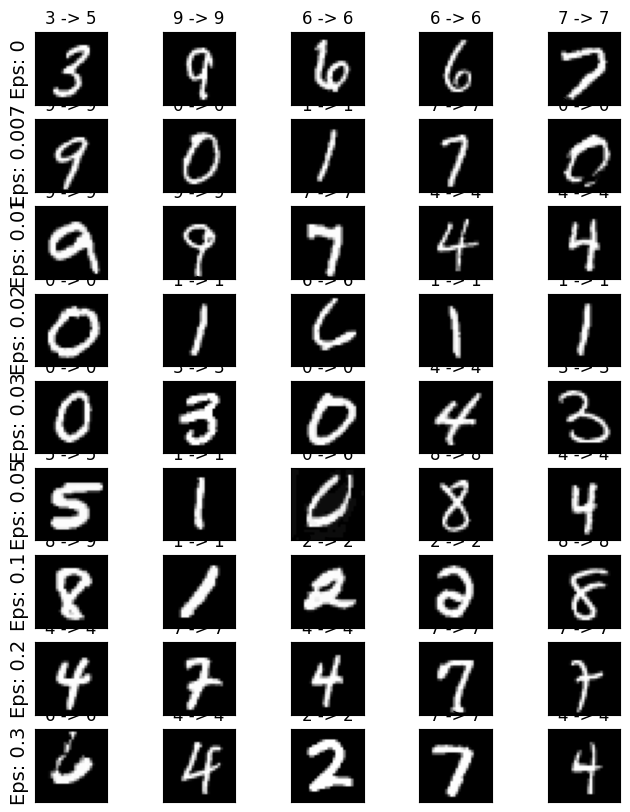


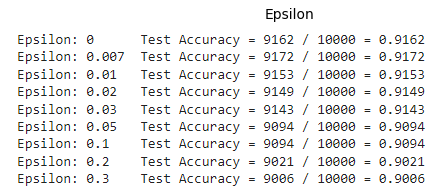
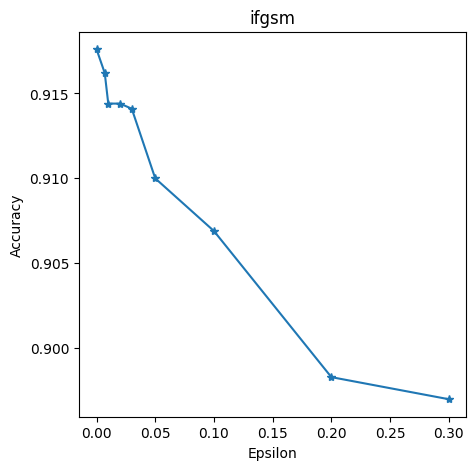


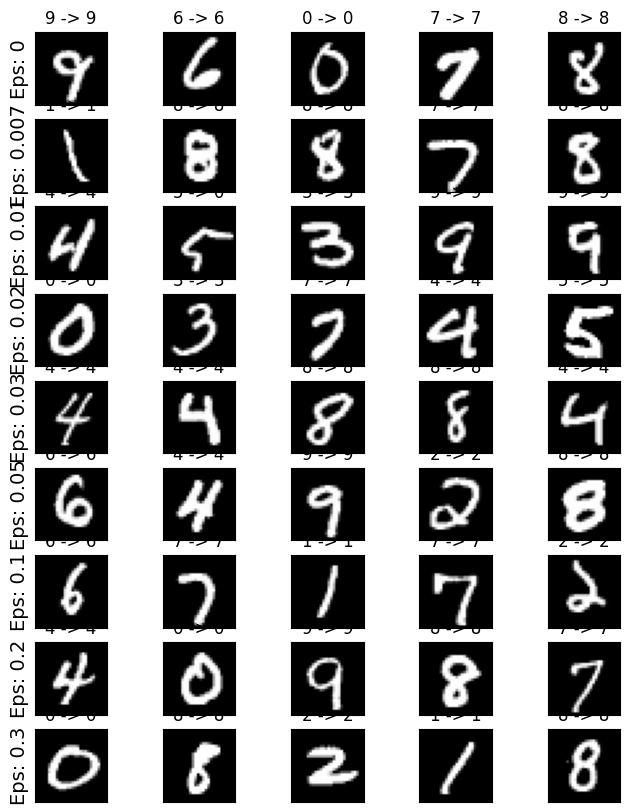


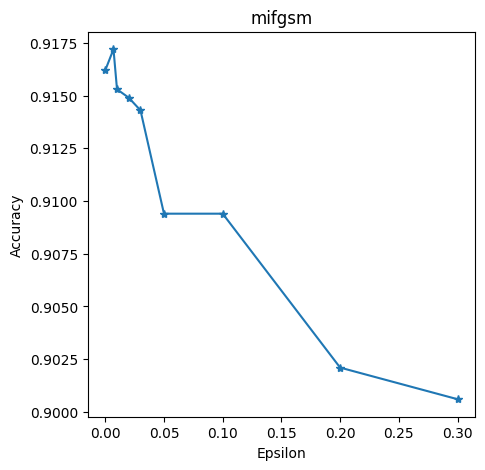


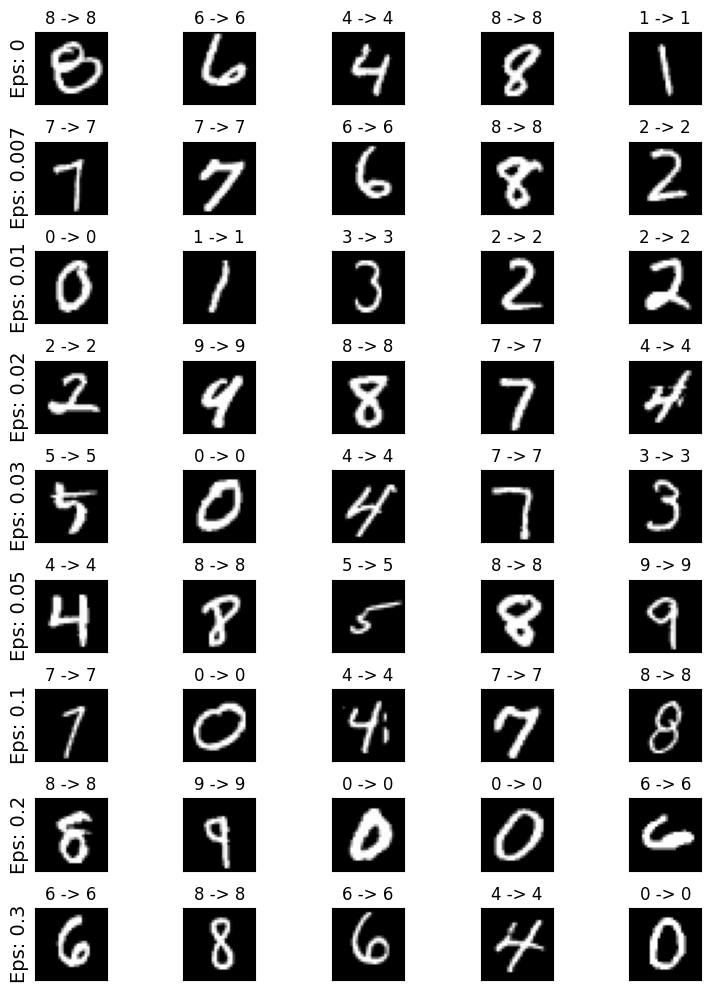












17. Сделать выводы по полученным результатам, оценить стойкость предложенного способа защиты моделей НС

Защитная дистилляция (Defensive Distillation) является методом защиты модели машинного обучения от атак, основанных на использовании обратных атак или атак переноса. Этот метод был предложен в 2016 году Филемми Шайи и Юрием Янгом.

Основная идея защитной дистилляции заключается в повторном обучении модели с использованием двух процессов: дистилляции и обучения защитного классификатора.

В процессе дистилляции:

1. Исходная модель, называемая учителем, обучается на исходных данных.

2. Затем полученные предсказания учителя используются для учебника, построенного на основе этих предсказаний. Учебник является более компактной и менее сложной моделью, называемой учеником.

3. Ученик затем обучается всему набору данных с учителем, используя предсказания учителя в качестве дополнительной информации.

В процессе обучения защитного классификатора:

1. Используется учебник, полученный из первого шага дистилляции, а также оригинальный набор данных исходной модели.

2. Защитный классификатор обучается на этом наборе данных, используя учебник как дополнительный вход для повышения устойчивости модели к атакам.

Цель защитной дистилляции заключается в том, чтобы сделать модель более устойчивой к атакам, например, к обратному инжинирингу или к атакам переноса, обеспечивая стройную и компактную модель, которая может сохранить высокую производительность.

Итог по увеличению стойкости модели:

атака fgsm снизила точность для обычной модели до - 22%, для защищенной - до - 90%;

атака ifgsm снизила точность не защищенных данных до - 25%, защищенных - до - 90%; атака

mifgsm снизила точность не защищенных данных до - 24%, защищенных - до - 90%.