

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

**«МИРЭА – Российский технологический университет»**

РТУ МИРЭА

ИКБ направление «Киберразведка и противодействие угрозам с применением технологий искусственного интеллекта» 10.04.01

Кафедра КБ-4 «Интеллектуальные системы информационной

безопасности»

Практическая работа №6

по дисциплине

«Анализ защищенности систем искусственного интеллекта»

Группа: ББМО-02-22

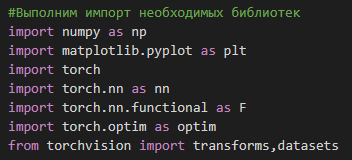
Выполнил:

Поплавский М.С.

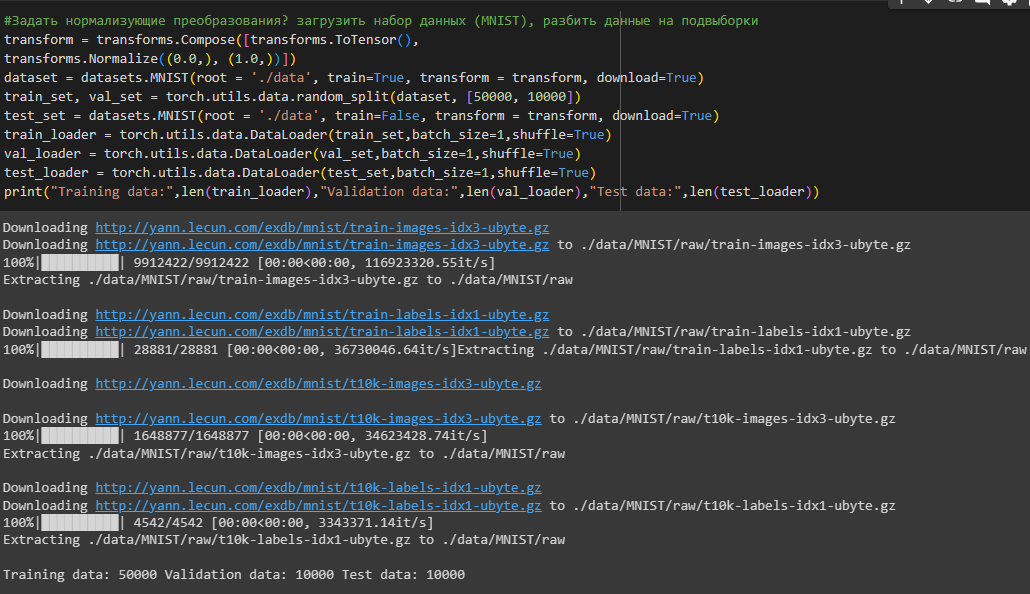
Проверил: Спирин А.А.

Москва 2023

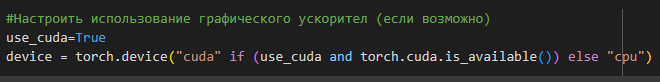
Выполняем импорт необходимых библиотек



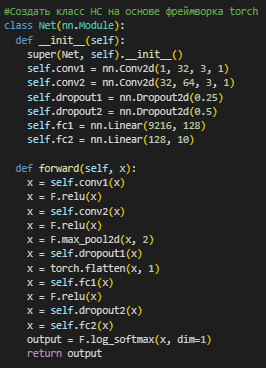
Задаем нормализующие преобразования, загружаем набор MNIST и разбиваем данные на подвыборки.



Задаем использование GPU для вычислений, если это возможно.

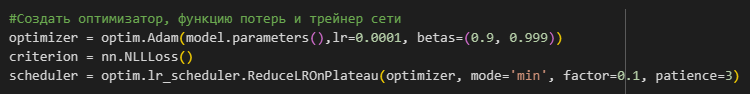


На основе torch создаем класс нейронной сети

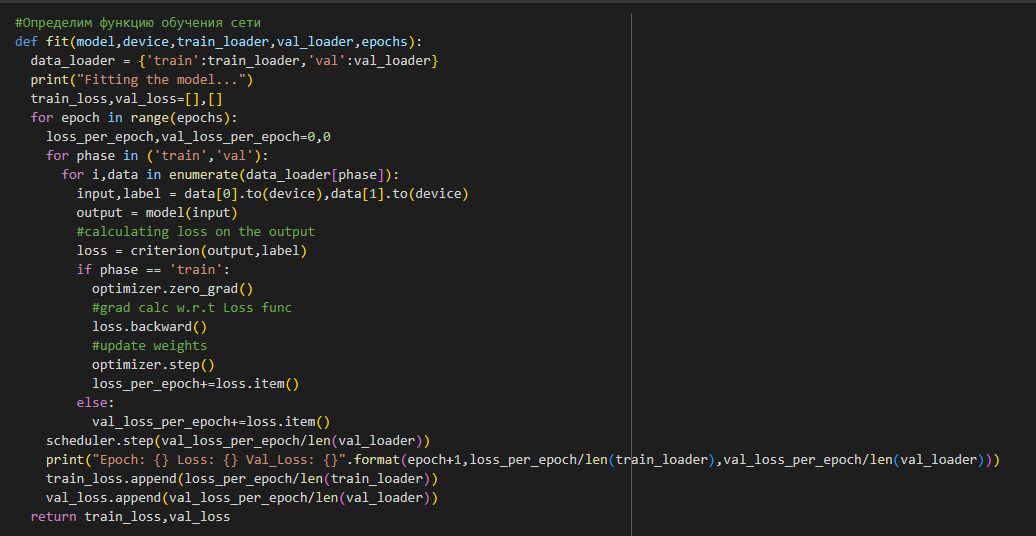


Проверяем работоспособность созданного класса и создаем оптимизатор, функцию потерь и трейнер сети.

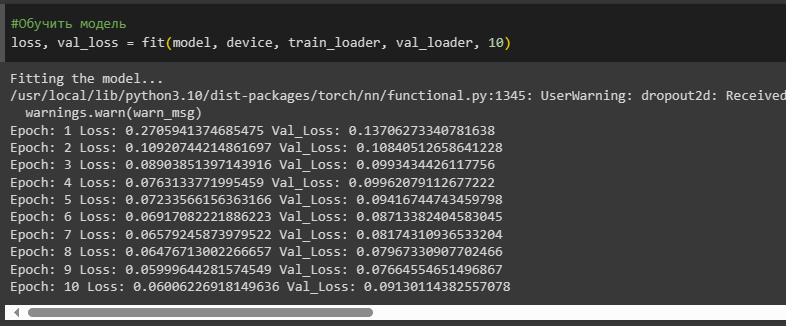




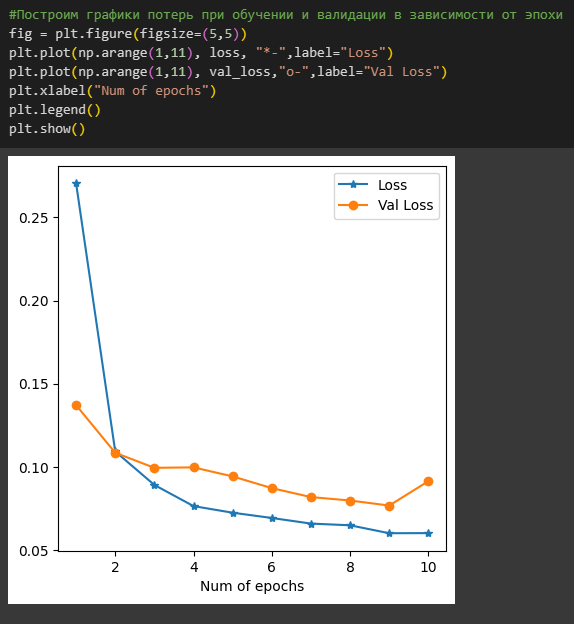
Определяем функцию обучения сети.



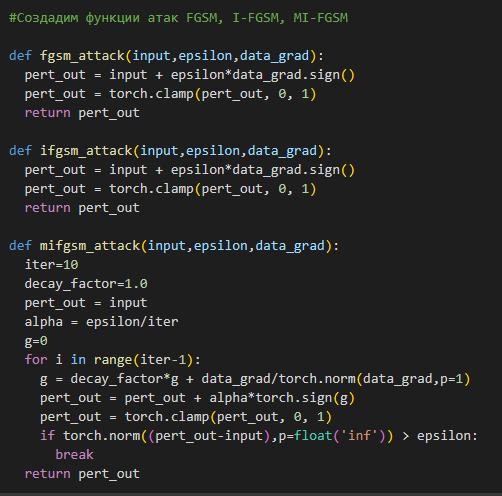
Выполним обучение модели.



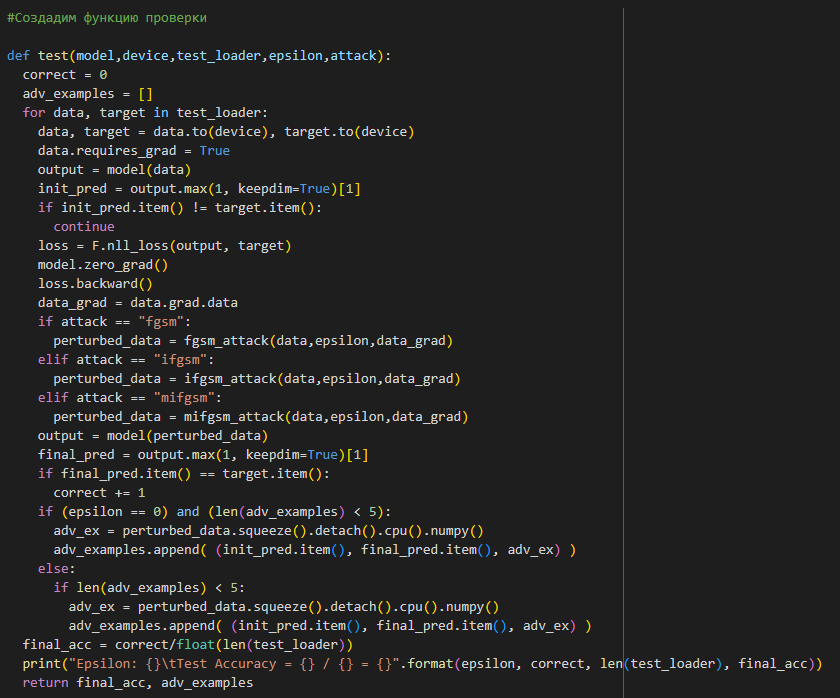
Строим график



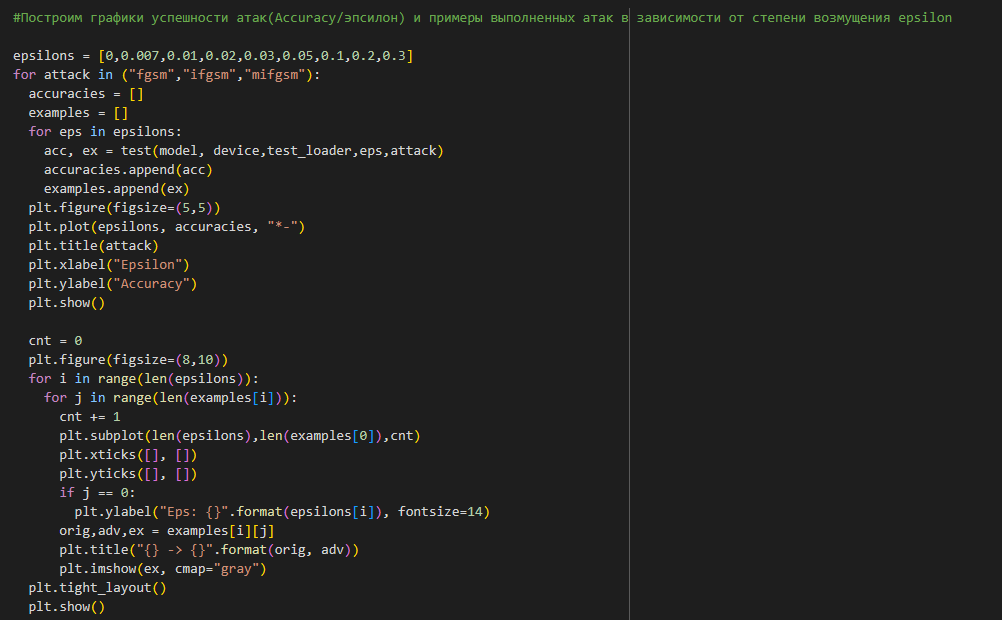
Создаем функции атак

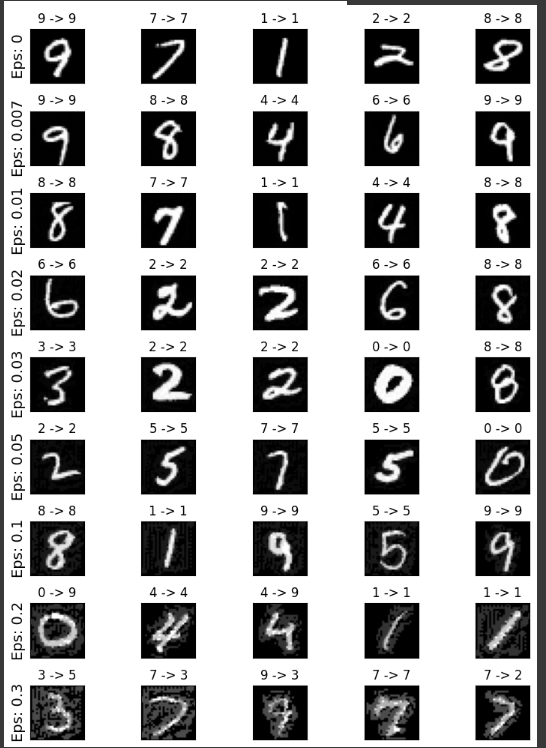
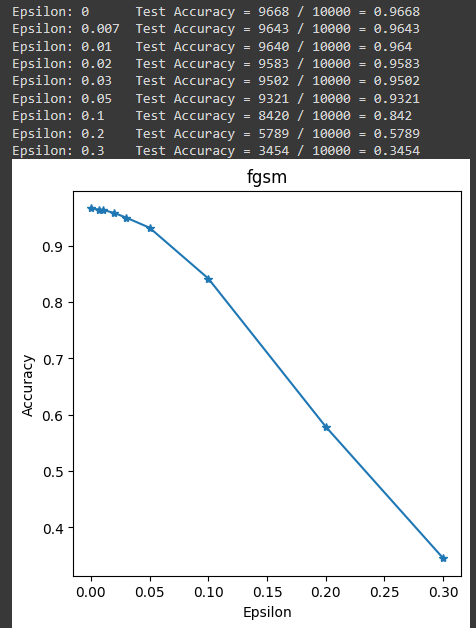


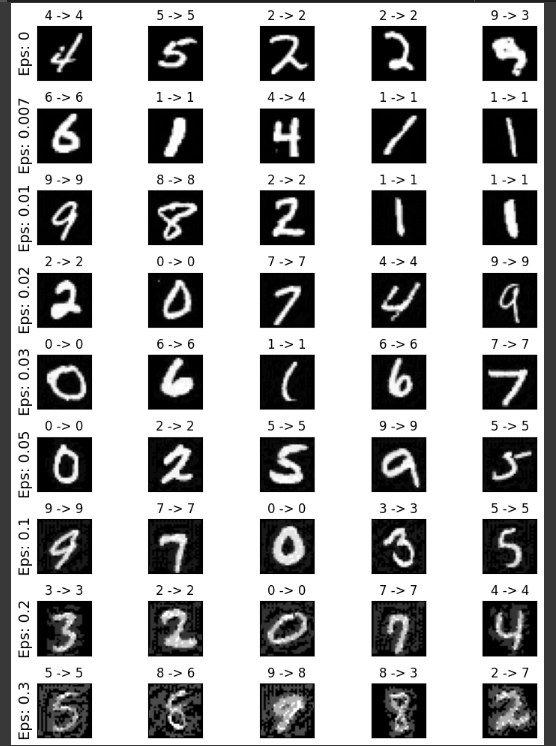
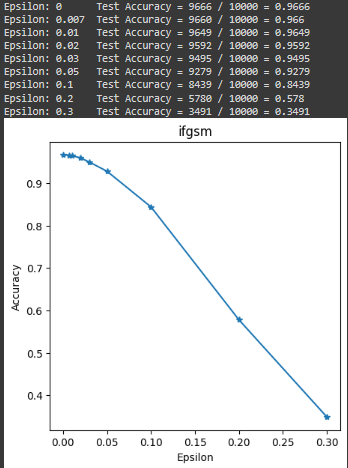
Создаем функцию для проверки.

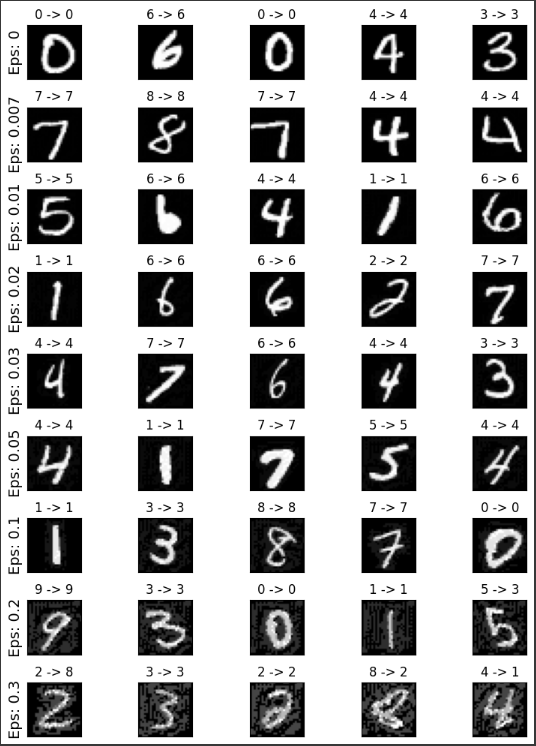
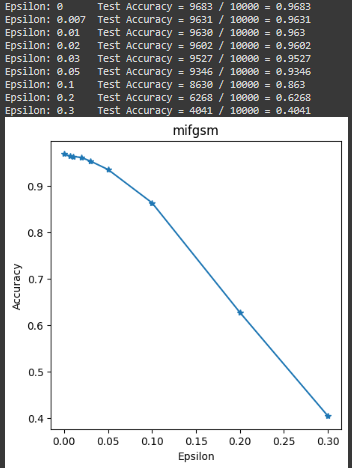


Выполним построение графиков успешности атак и примеров выполняемых атак в зависимости от степени возмущения epsilon

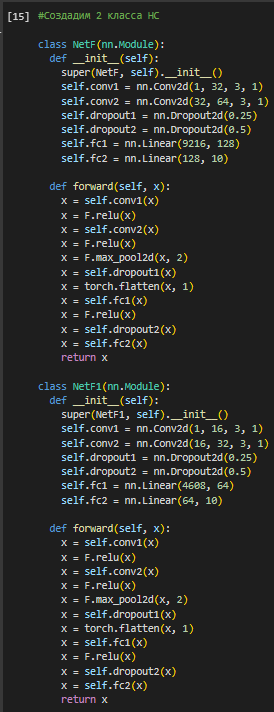




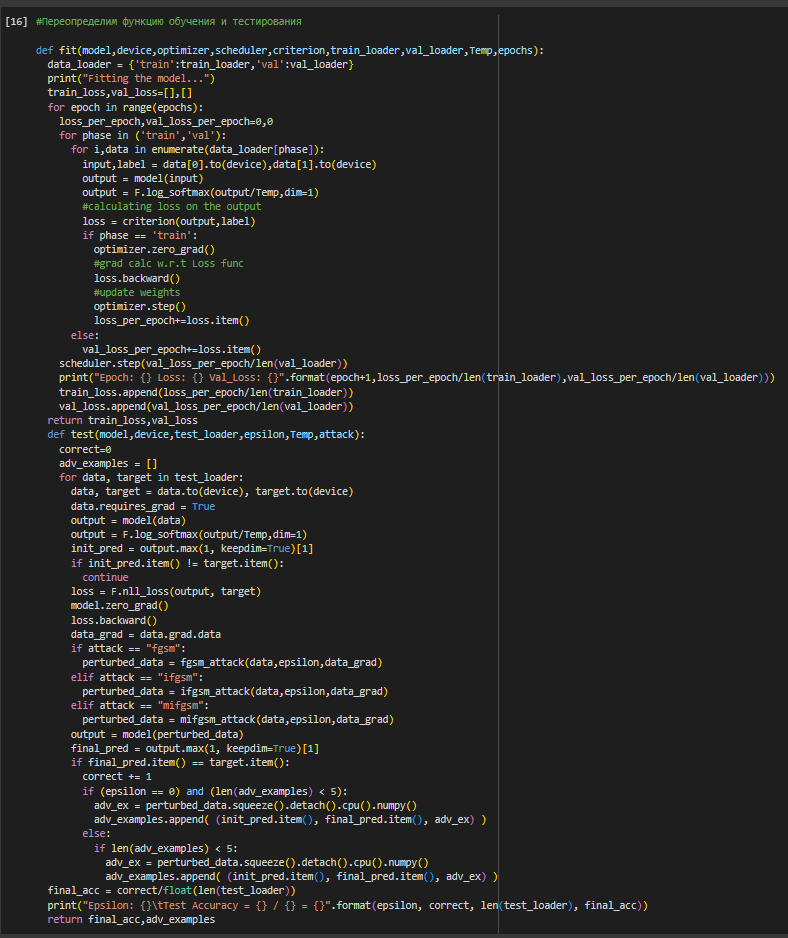




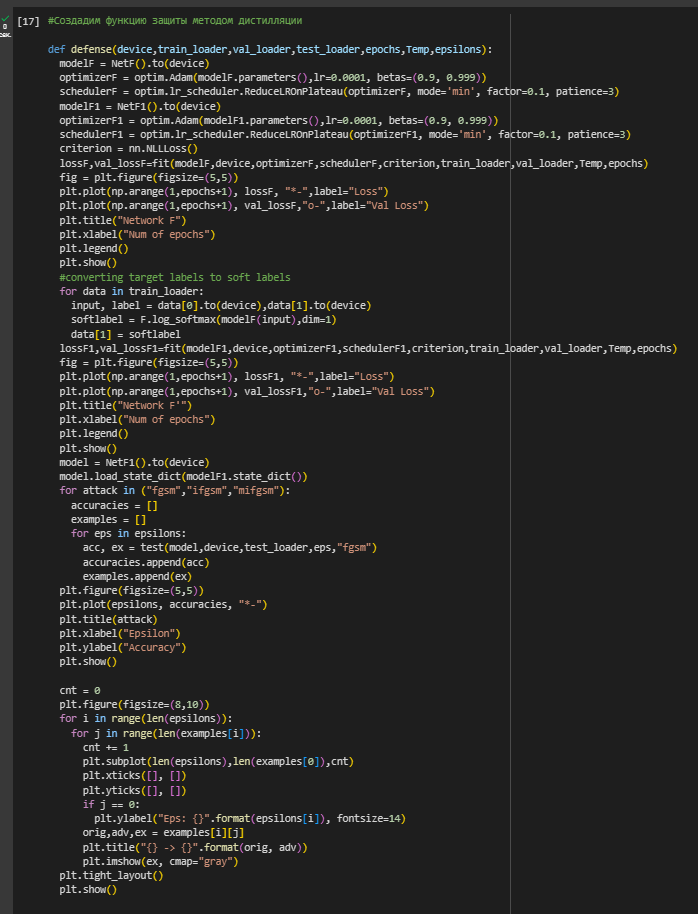
Создадим два класса нейросети



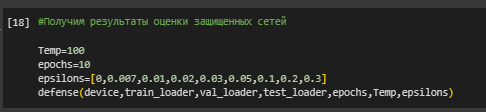
Создадим функции обучения и тестирования

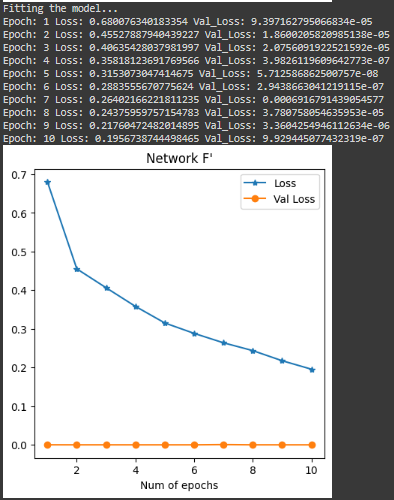
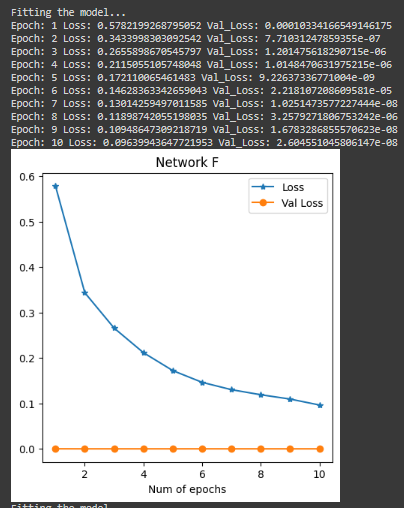


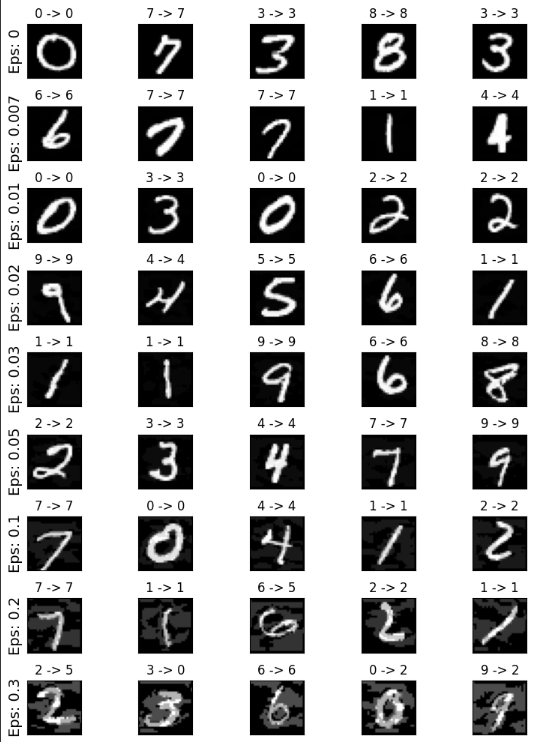
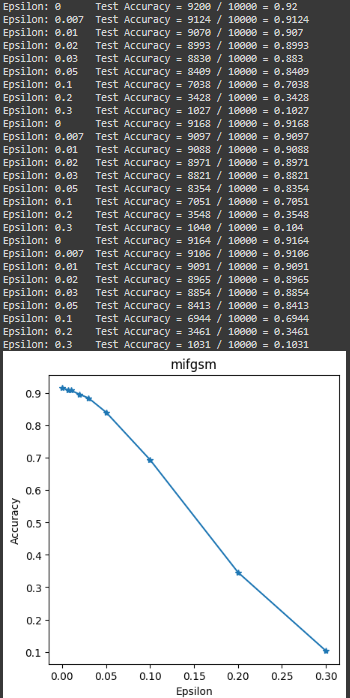
Зададим функцию защиты сетей.



Получим результаты оценки защищенных сетей







**Вывод.**

В данной работе мы успешно создали модели и загрузили данные, обучили их, протестировали и применили защитную дистилляцию, которая представляет собой метод защиты нейронных сетей от атак, направленных на восстановление информации из параметров.

Без использования защитных механизмов (при eps=0) точность модели высока, и она составляет 91.28%. С увеличением значения eps точность модели ухудшается, особенно при больших значениях. Это свидетельствует о том, что модель чувствительна к атакам с добавлением шума (например, FGSM, iFGSM, MI-FGSM).

Визуализация примеров атак позволяет увидеть, как изменяется внешний вид изображений после атаки для различных значений eps.

Можно заметить, что с увеличением eps изображения становятся все более искаженными и трудными для классификации моделью.

В данной работе учителем является NetF, учеником NetF1