

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

Институт комплексной безопасности и специального приборостроения

**Отчет** по лабораторной работе №4

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

# Выполнил:

Студент группы ББМО-01-22 ФИО: Карпова Н.Ю.

# Москва 2023

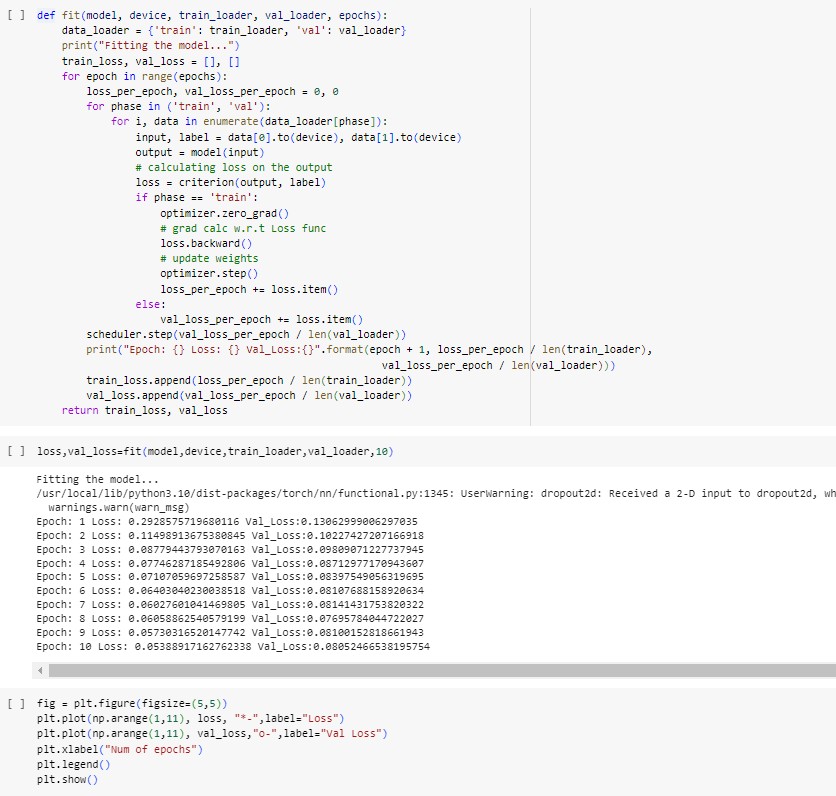
1. Выполним импорт библиотек и загрузим набор данных и настроим использование графического устройства.



1. Создадим класс НС на основе фреймворка torch и проверим работоспособность.



1. Определим функцию обучения и обучим модель. Построим график потерь.



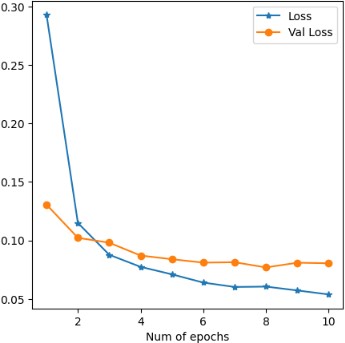
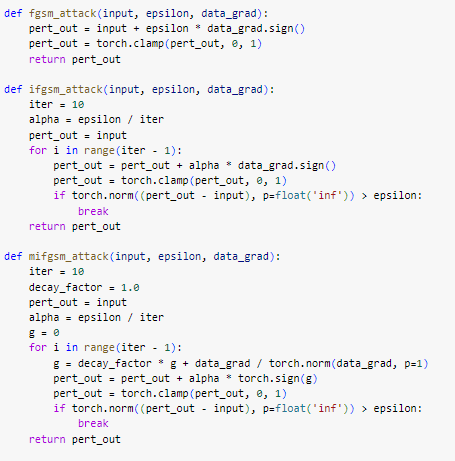
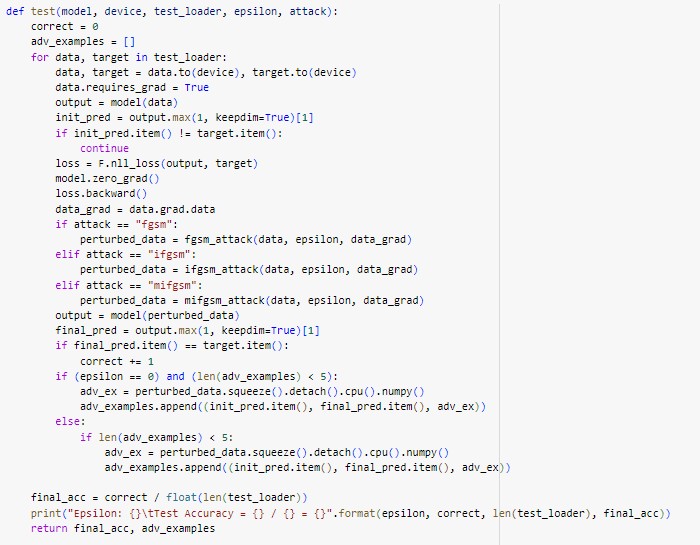


Рисунок 1 – График потерь при обучении и валидации в зависимости от эпохи

1. Создадим функции атак (FGSM, I-FGSM, MI-FGSM)



1. Создадим Функцию проверки



1. Воспроизведем график успешности атак и примеры выполных атак в зависимости от степени возмущения.



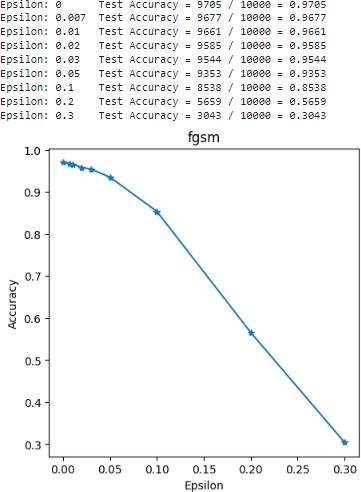


Рисунок 2 - Результат

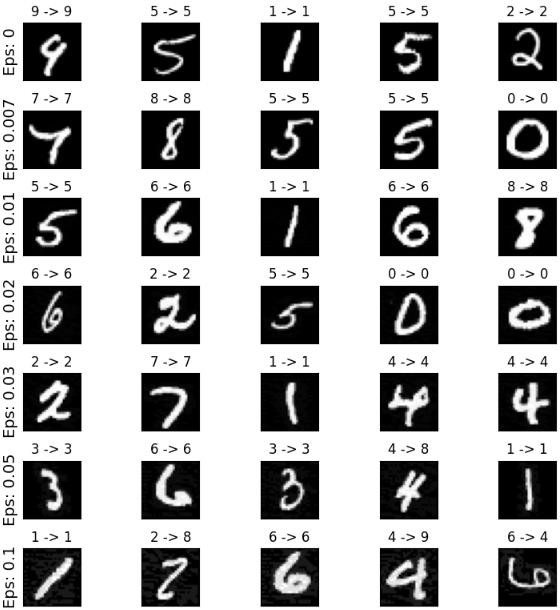


Рисунок 3 - Результат

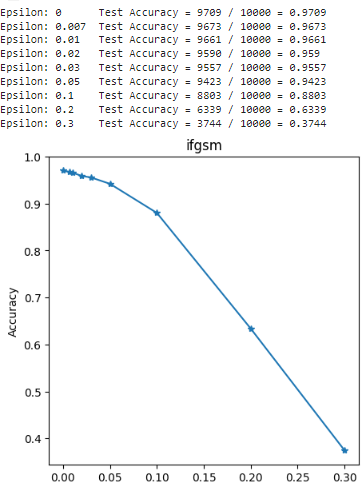


Рисунок 4 - Результат

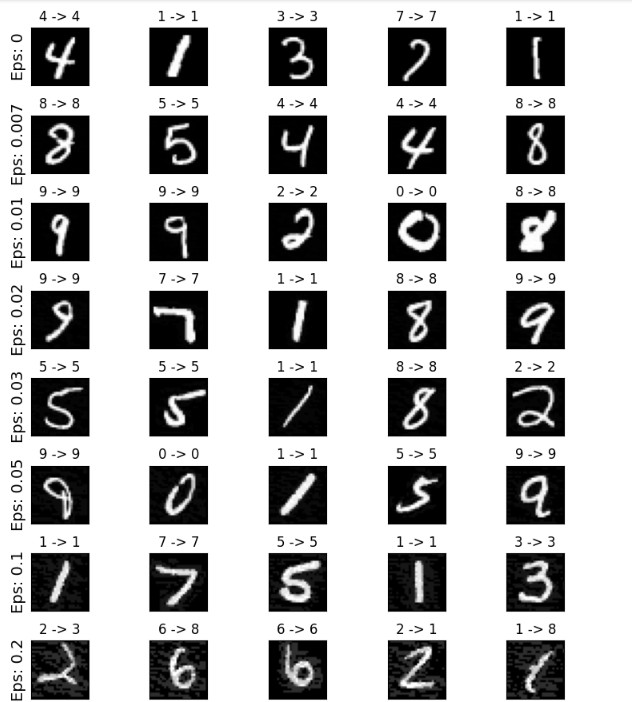


Рисунок 5 - Результат

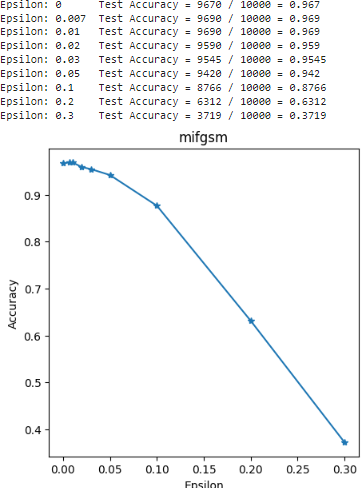


Рисунок 6 - Результат

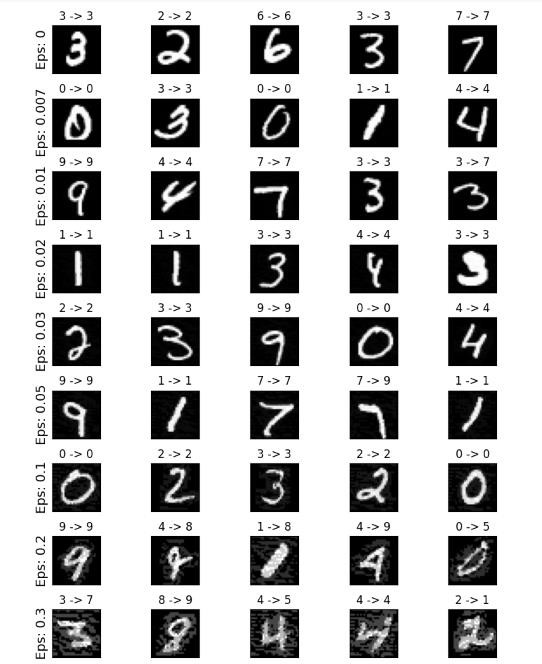
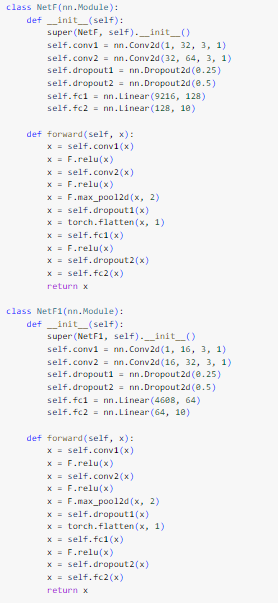
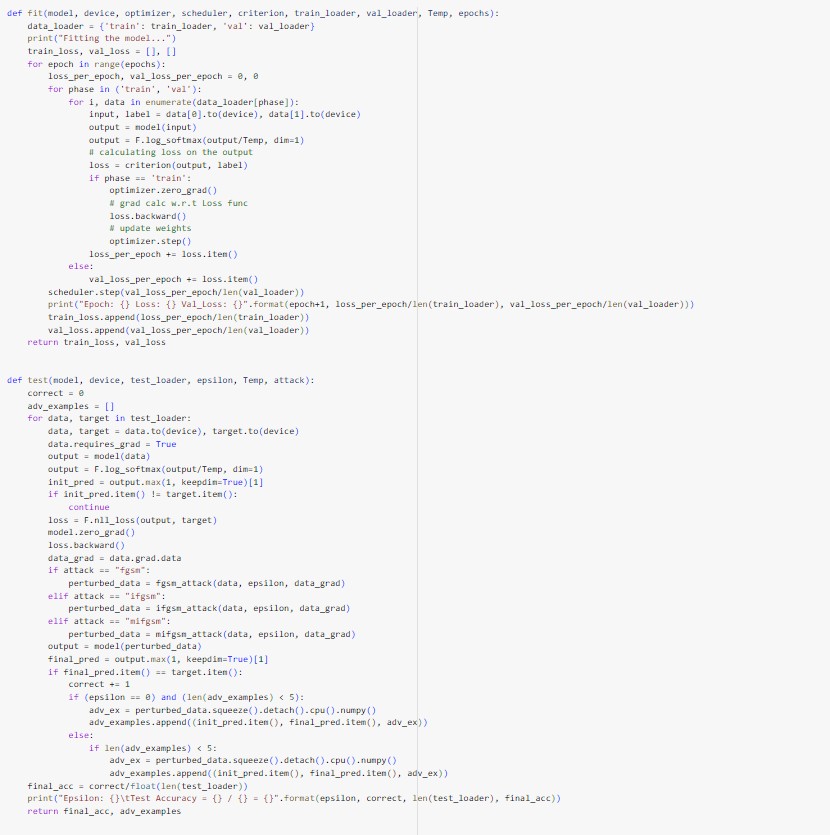


Рисунок 7 - Результат

1. Создадим 2 класса.



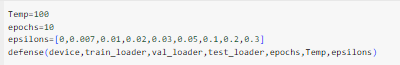
1. Преобразуем функцию обучения и тестирования.

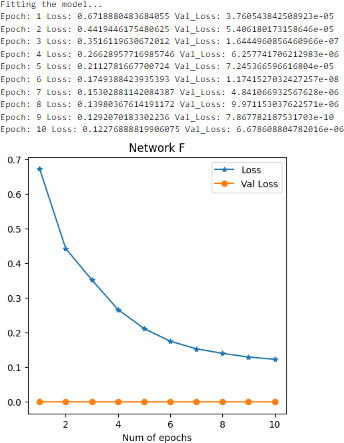


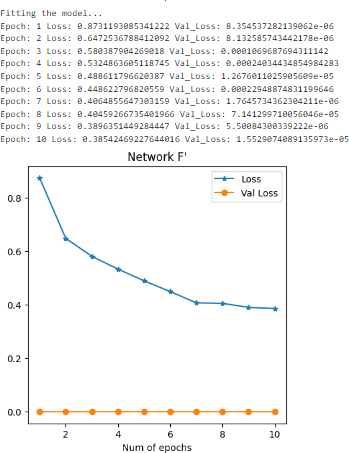
1. Сделаем функцию защиты методом дистилляции.

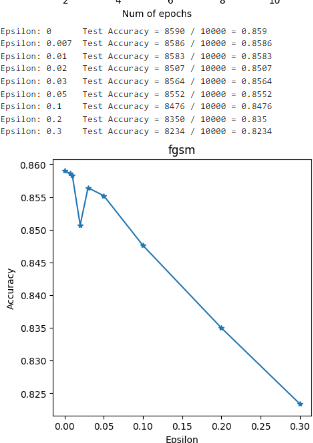


1. Получим результат оценки защищенных сетей

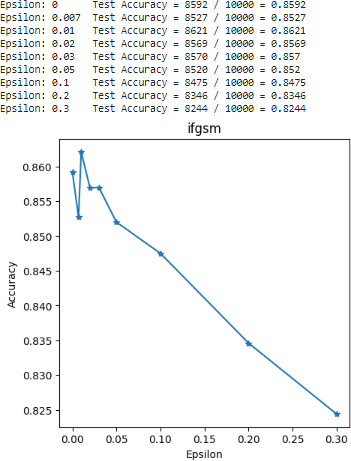


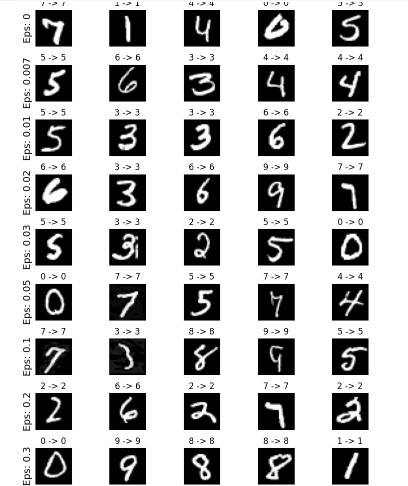


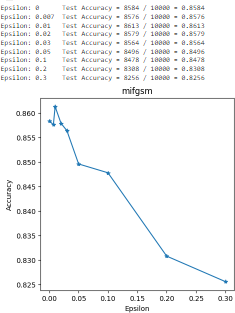


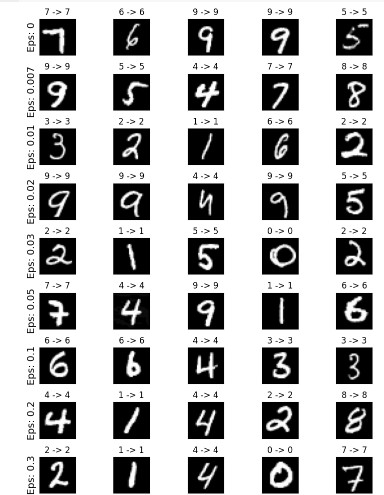












# Вывод

На основе выполнения лабораторной работы можно сделать следующие выводы:

Обучение с двумя моделями: Использование двух моделей (NetF и NetF1), где NetF1 получает мягкие метки от NetF, является стратегией дистилляции знаний. Это может помочь в передаче "опыта" более устойчивой модели NetF менее устойчивой модели NetF1. Графики потерь и потерь на валидационном наборе данных для обеих моделей позволяют оценить их обучение. Отслеживание изменений во времени помогает определить, происходит ли переобучение. Оценка точности модели NetF1 после обучения с использованием предложенного механизма защиты под воздействием различных атак (например, fgsm, ifgsm, mifgsm) при разных значениях эпсилон предоставляет информацию о ее устойчивости к атакам.