|  |  |  |
| --- | --- | --- |
|  |  |  |
| 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 №4**

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

Выполнил:

Евдокимов А.М.

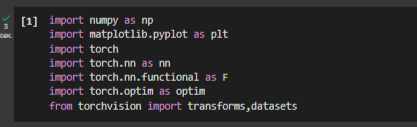
Группа: ББМО-02-23

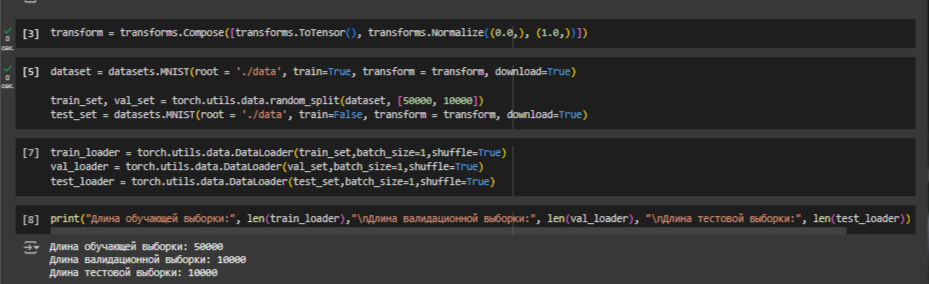
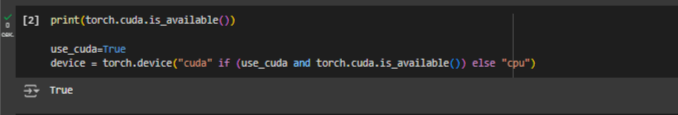
Москва - 2024

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

**Защитная дистилляция**

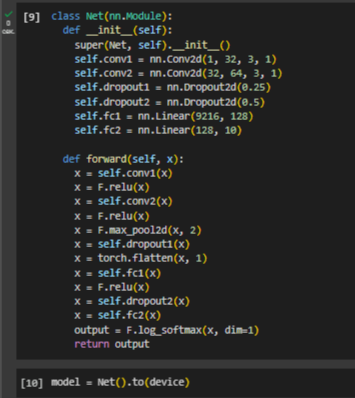
1. Выполнить импорт необходимых библиотек.



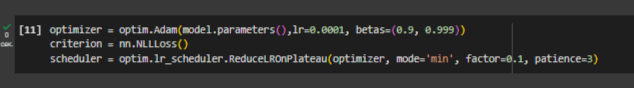
1. Загрузим набор данных (MNIST), разобьем данные на подвыборки
2. Настроим использование графического ускорителя 

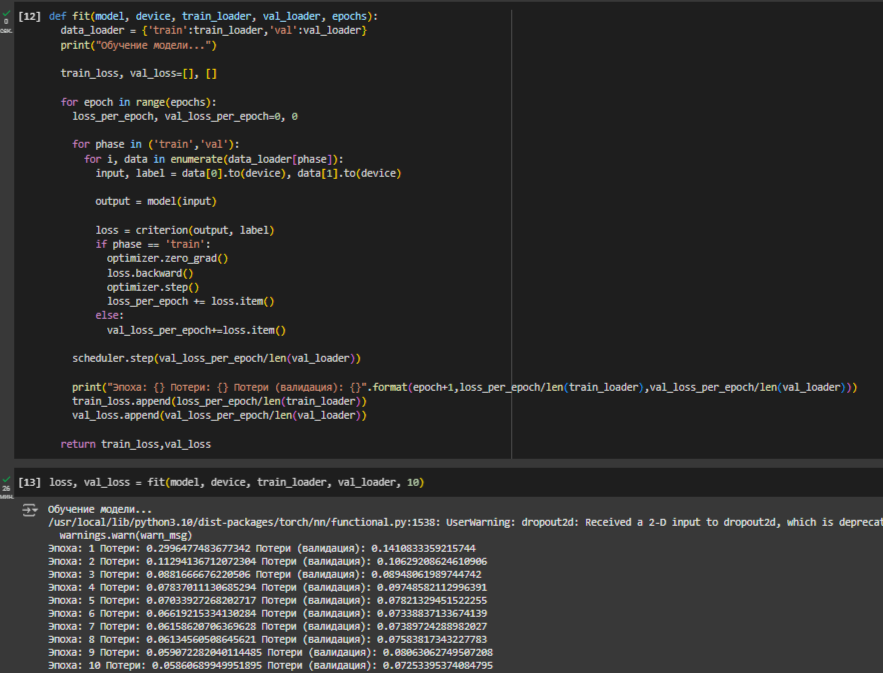
**Создание атак на модель НС**

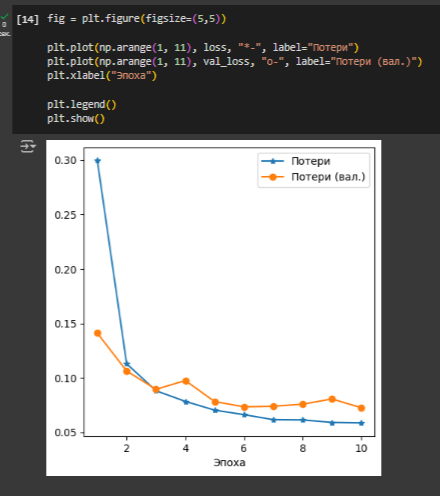
1. Создать класс НС на основе фреймворка torch и проверяем работоспособность созданного класса НС

****

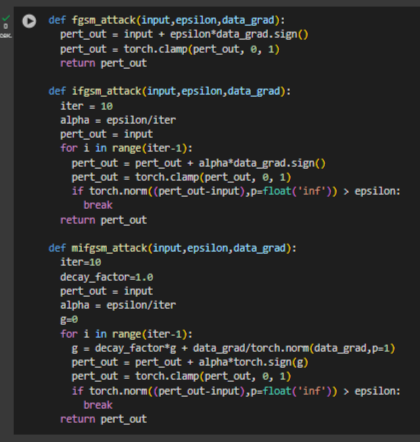
1. Создать оптимизатор, функцию потерь и трейнер сети



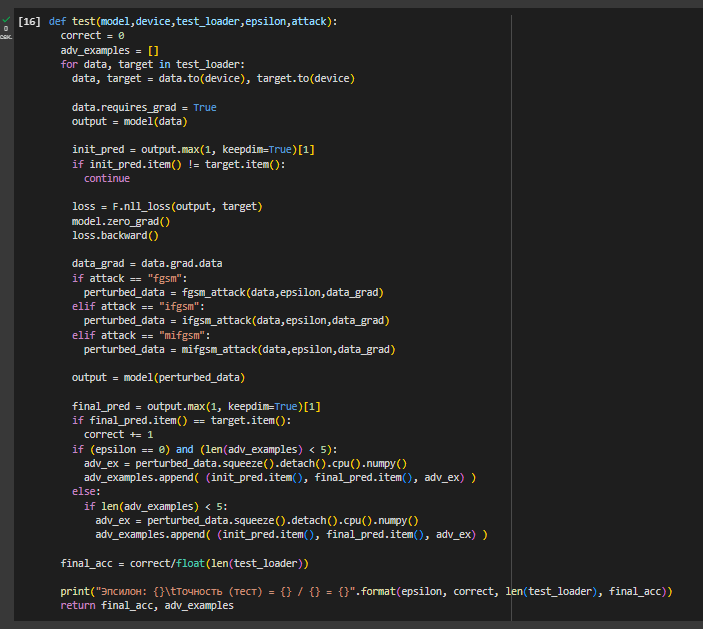
1. Определить функцию обучения сети и обучаем модель
2. Построим графики потерь при обучении и валидации в зависимости от эпохи



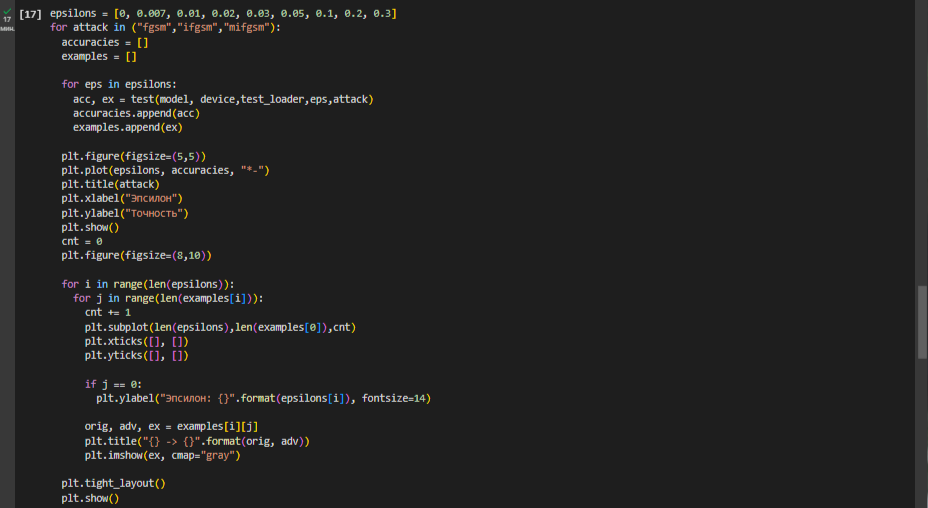
1. Создадим функции атак FGSM, I-FGSM, MI-FGSM

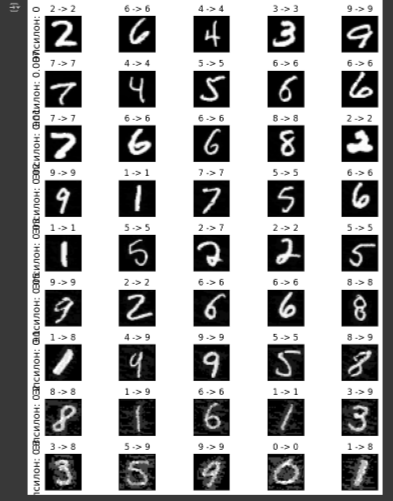


1. Создадим функцию проверки



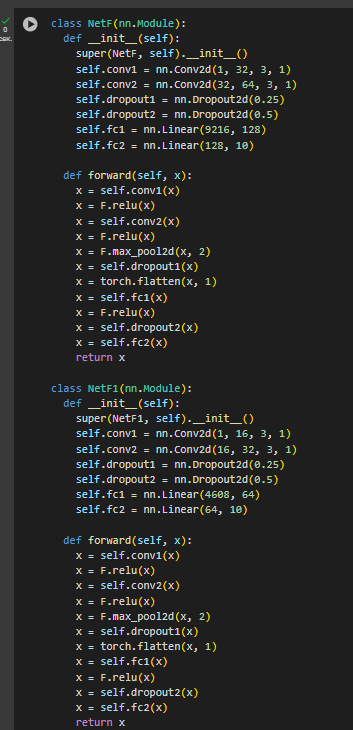
1. Построим графики успешности атак(Accuracy/эпсилон) и примеры выполненных атак в зависимости от степени возмущения epsilon:



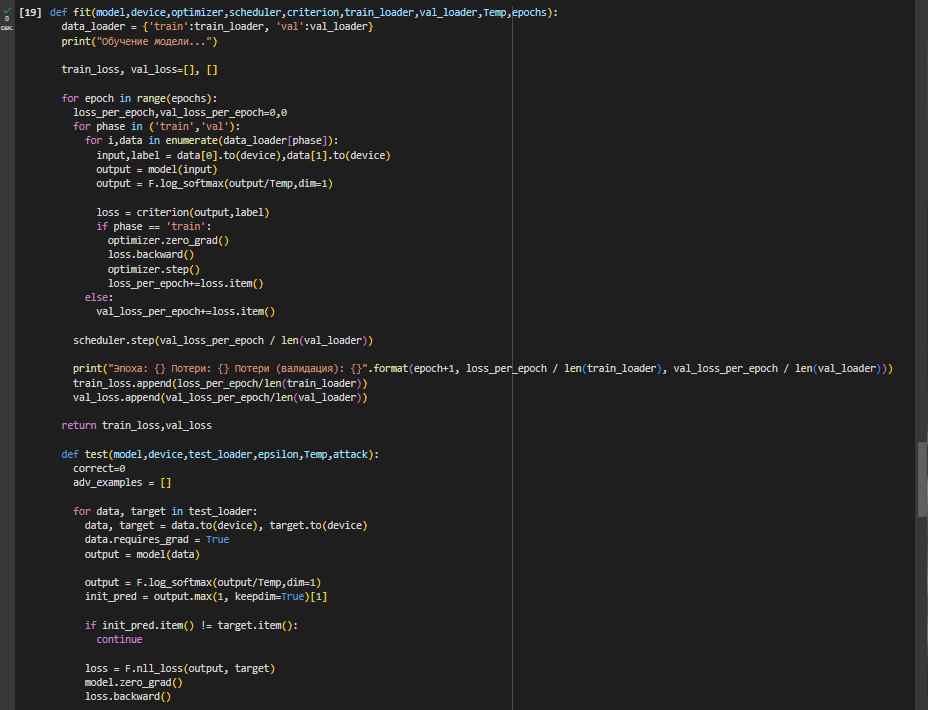


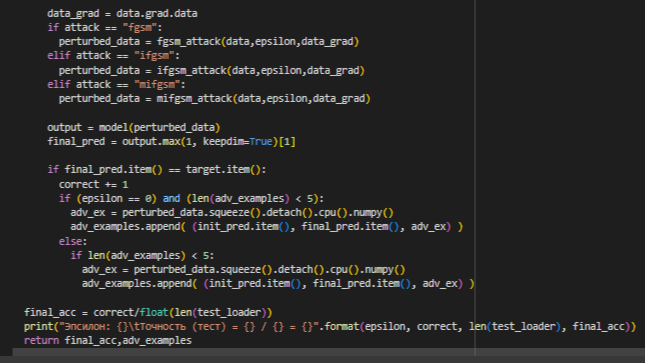
**Защита от атак**

1. Создадим 2 класса НС

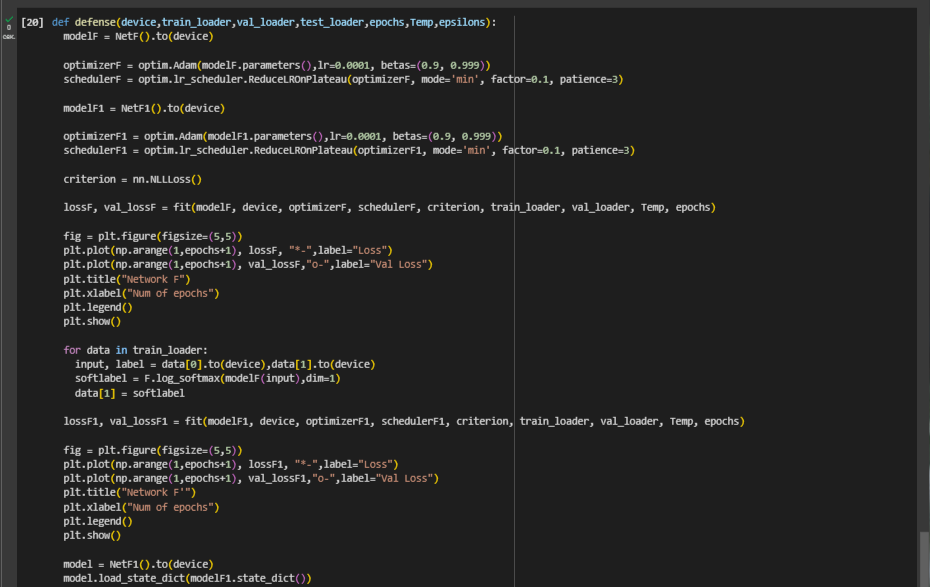
****

1. Переопределим функцию обучения и тестирования

****

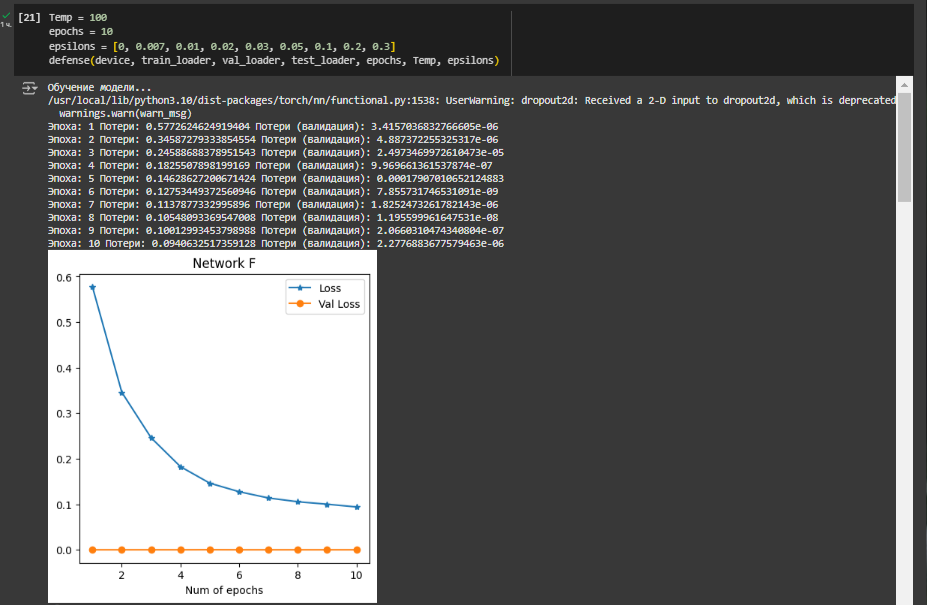
****

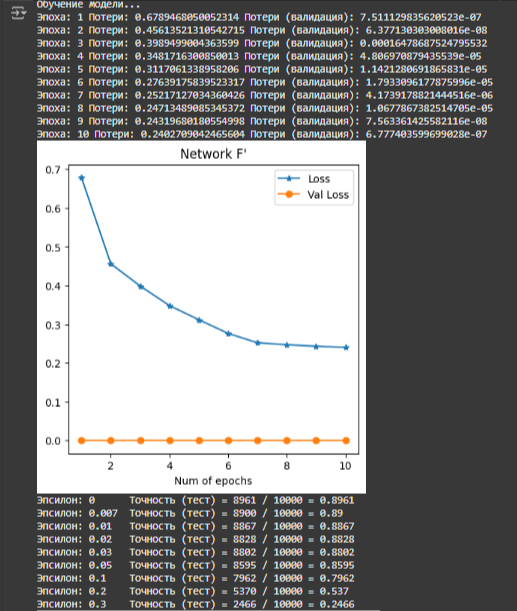
1. Создадим функцию защиты методом дистилляции

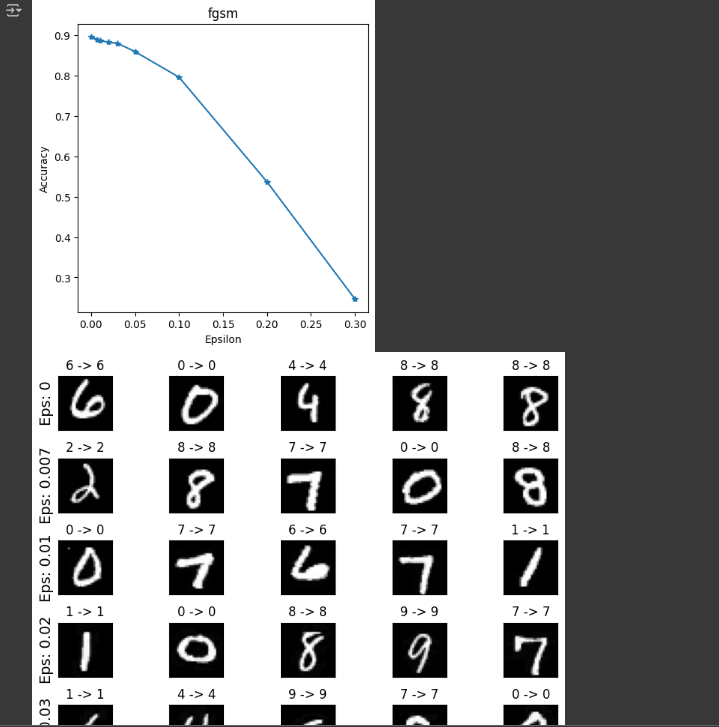
****

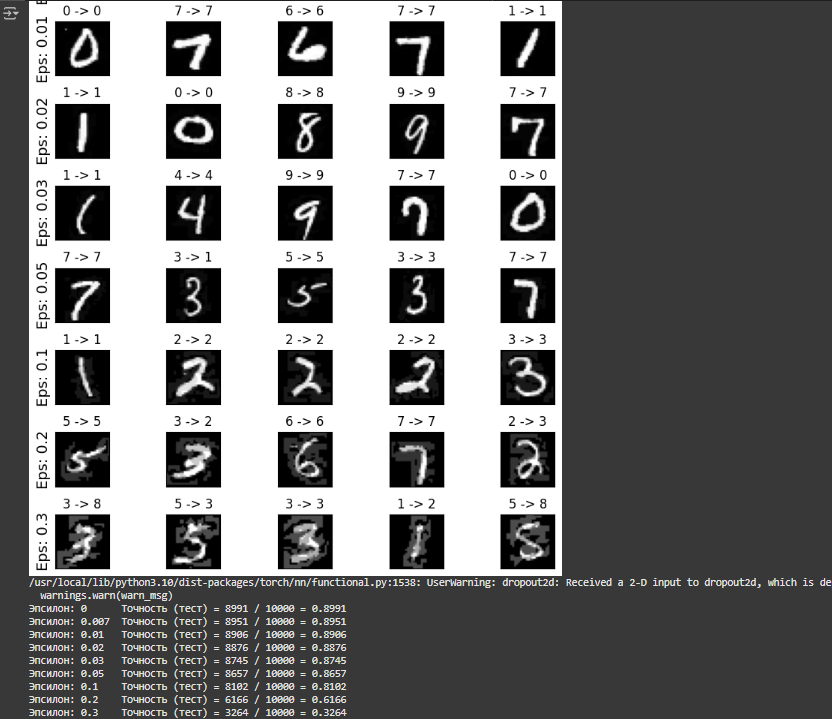
****

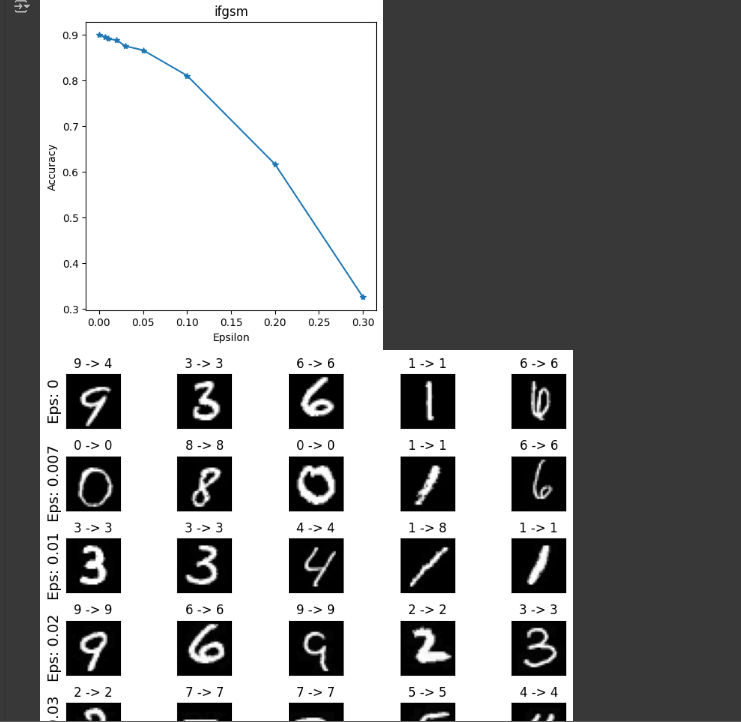
1. Получим результаты оценки защищенных сетей

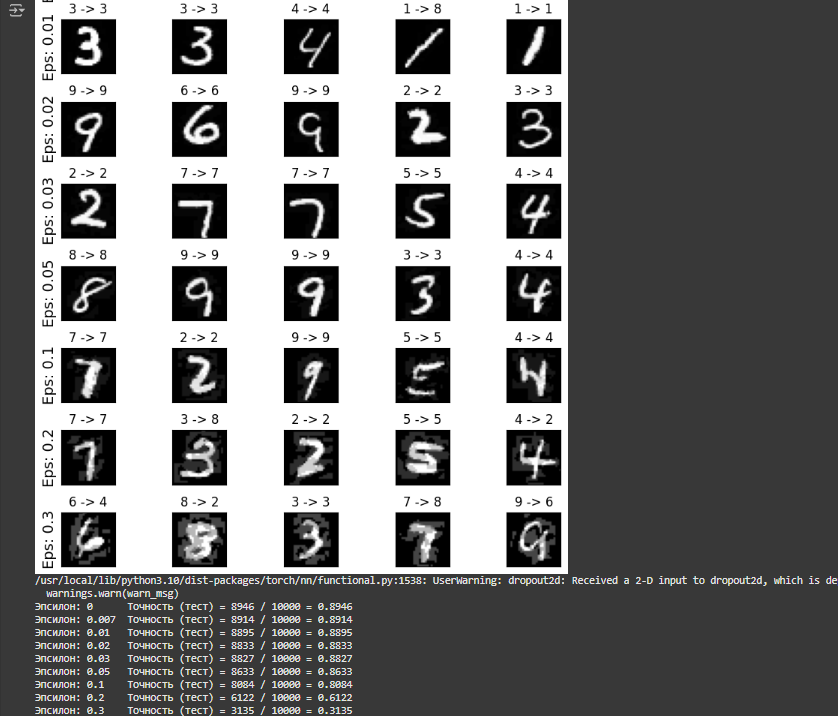


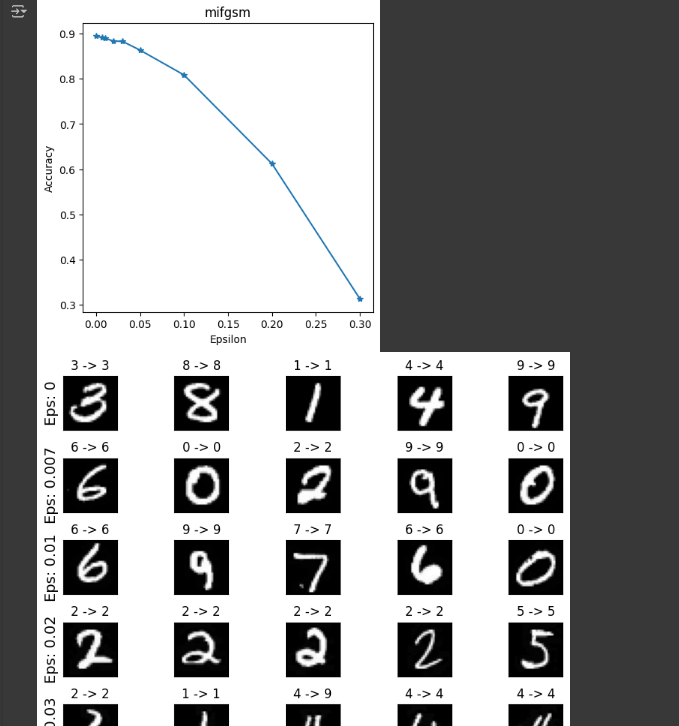














**Заключение**

Атаки FGSM, I-FGSM и MI-FGSM демонстрируют высокую эффективность при увеличении степени возмущения (ϵ).

Точность моделей снижается с ростом ϵ от почти 97% (без возмущения) до менее 20 - 25% при ϵ = 0.3, что свидетельствует о значительной уязвимости моделей без защиты.

Метод защитной дистилляции продемонстрировал повышение стойкости моделей к атакам:

Точность моделей при высоких значениях ϵ осталась выше, чем у моделей без защиты.

Защитная дистилляция уменьшила градиентные изменения, которые используются в атаках для создания возмущений.

MI-FGSM и I-FGSM показали немного большую стойкость моделей на уровне ϵ = 0.3 по сравнению с FGSM, но в целом различия минимальны.

Точность при низких значениях ϵ (0,007 - 0,03) осталась на приемлемом уровне для всех типов атак, но тенденция к снижению наблюдалась во всех случаях.

Метод защитной дистилляции эффективен для снижения успешности атак FGSM, I-FGSM и MI-FGSM, что подтверждается результатами тестов.

Простота интеграции метода делает его пригодным для использования в широком спектре задач.

При высоких значениях ϵ (>0,2) точность модели всё ещё существенно снижается, что говорит о необходимости дополнительных методов защиты для таких случаев.

Метод может быть менее эффективным против адаптивных атак, когда злоумышленники подстраиваются под защитные механизмы.

Защитная дистилляция доказала свою эффективность в качестве базового метода повышения стойкости моделей НС к атакам.

Для максимальной защиты рекомендуется сочетать данный подход с другими методами (например, шумоподавлением, регуляризацией или аугментацией данных) и тестировать модели против более сложных сценариев атак.