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

Институт кибербезопасности и цифровых технологий

КБ-4 «Интеллектуальные системы информационной безопасности»

Отчет по лабораторной работе №4

по дисциплине: «Анализ защищенности систем искусственного интеллекта»

**Выполнил**:

Студент группы ББМО-02-22

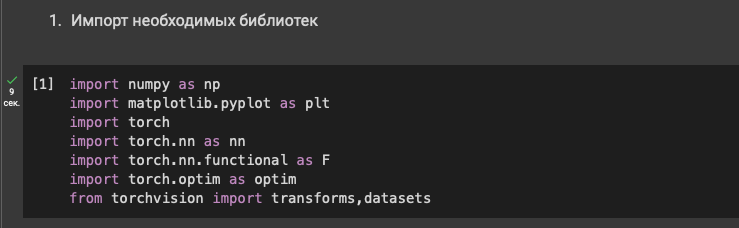
Филиппов Леонид Алексеевич

**Проверил**:

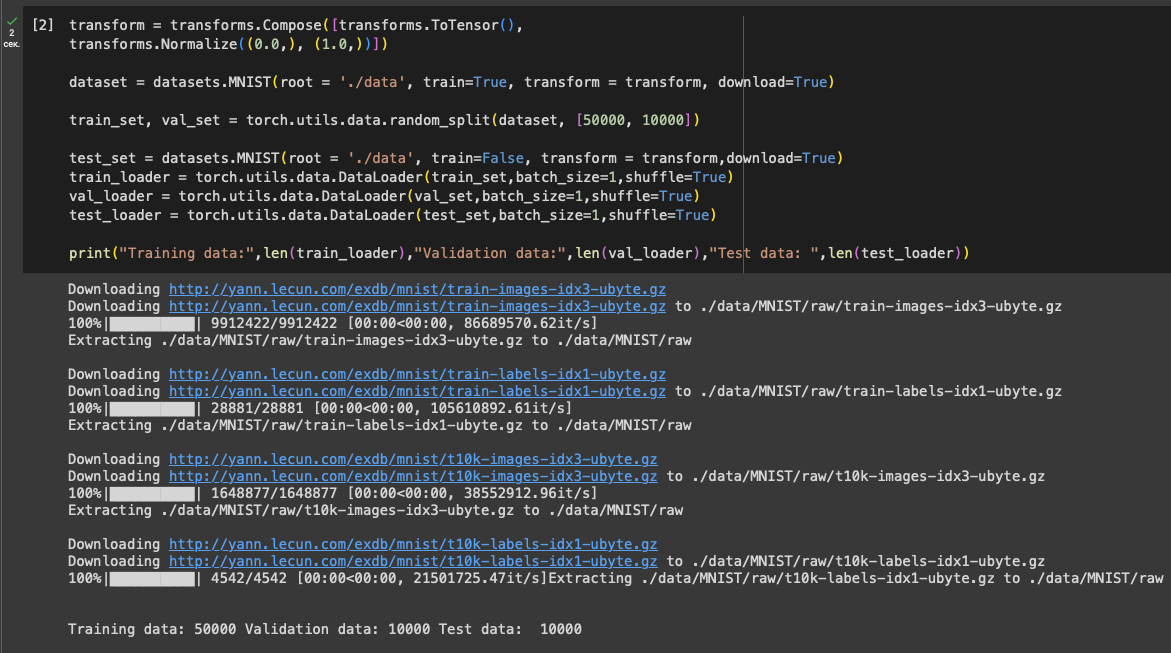
Спирин Андрей Андреевич

Москва 2024

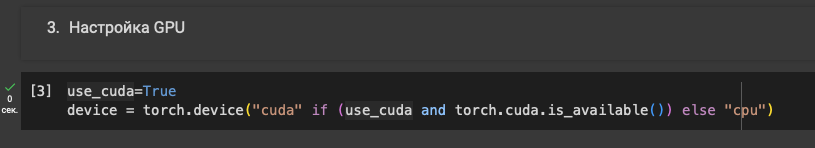
1. Импорт необходимых библиотек



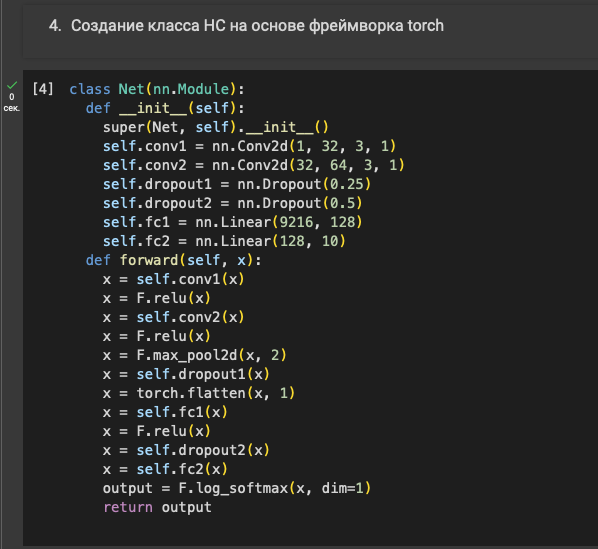
1. Задание нормализующих преобразований и загрузка набора данных (MNIST), разбиение данных на подвыборки



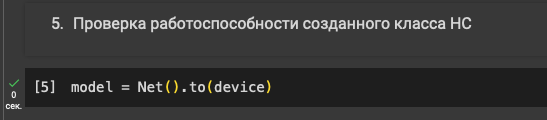
1. Настройка GPU



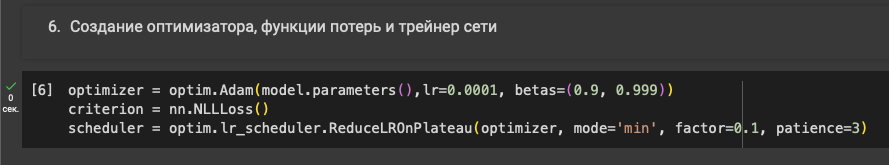
1. Создание класса НС на основе фреймворка torch



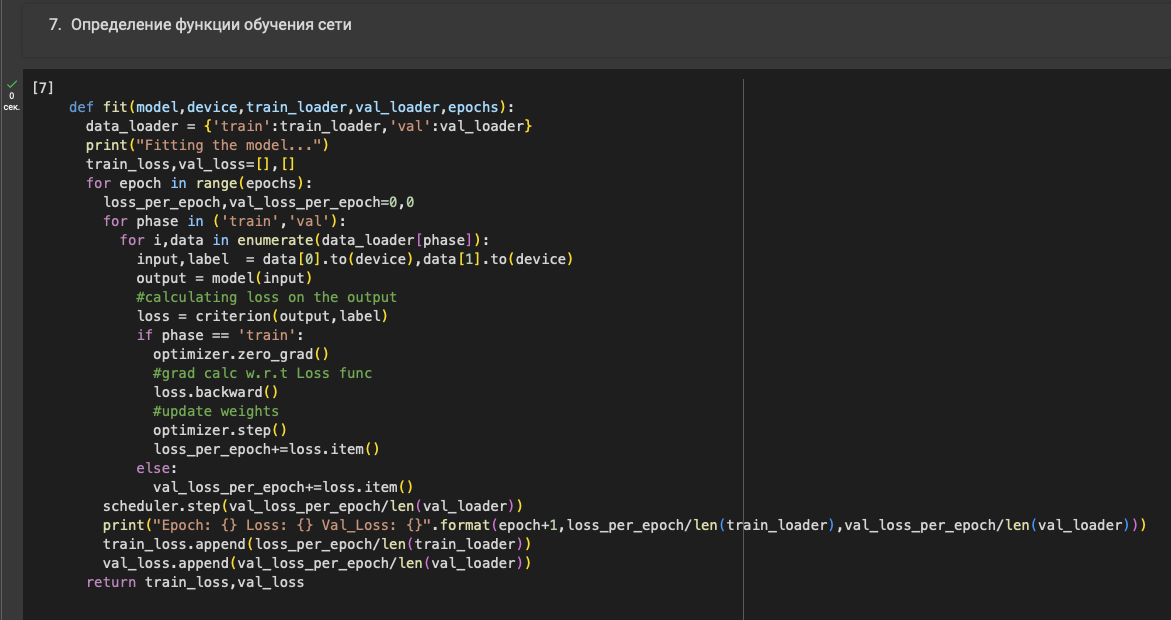
1. Проверка работоспособности созданного класса НС



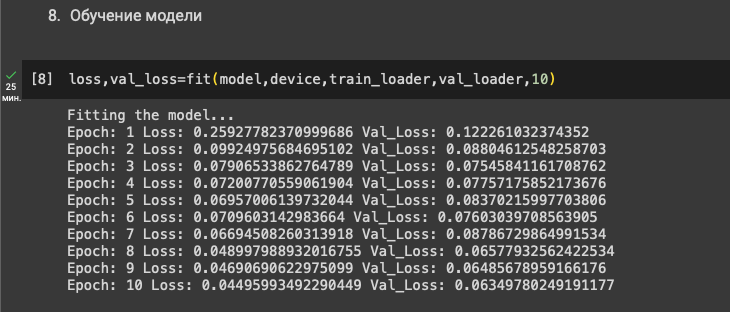
1. Создание оптимизатора, функции потерь и трейнер сети



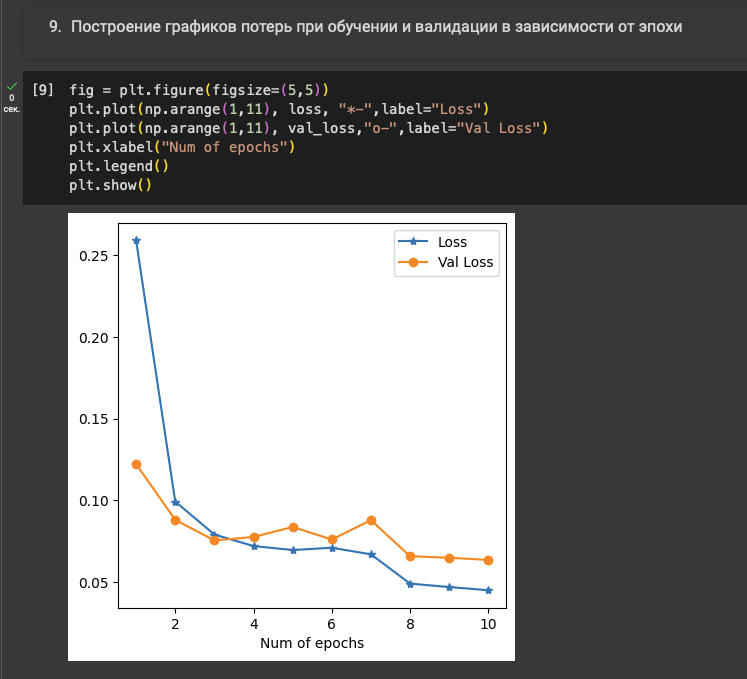
1. Определение функции обучения сети



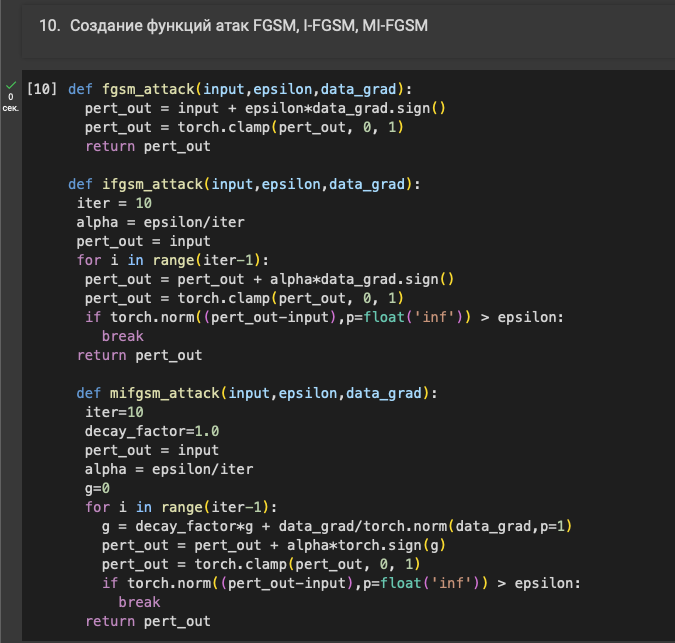
1. Обучение модели



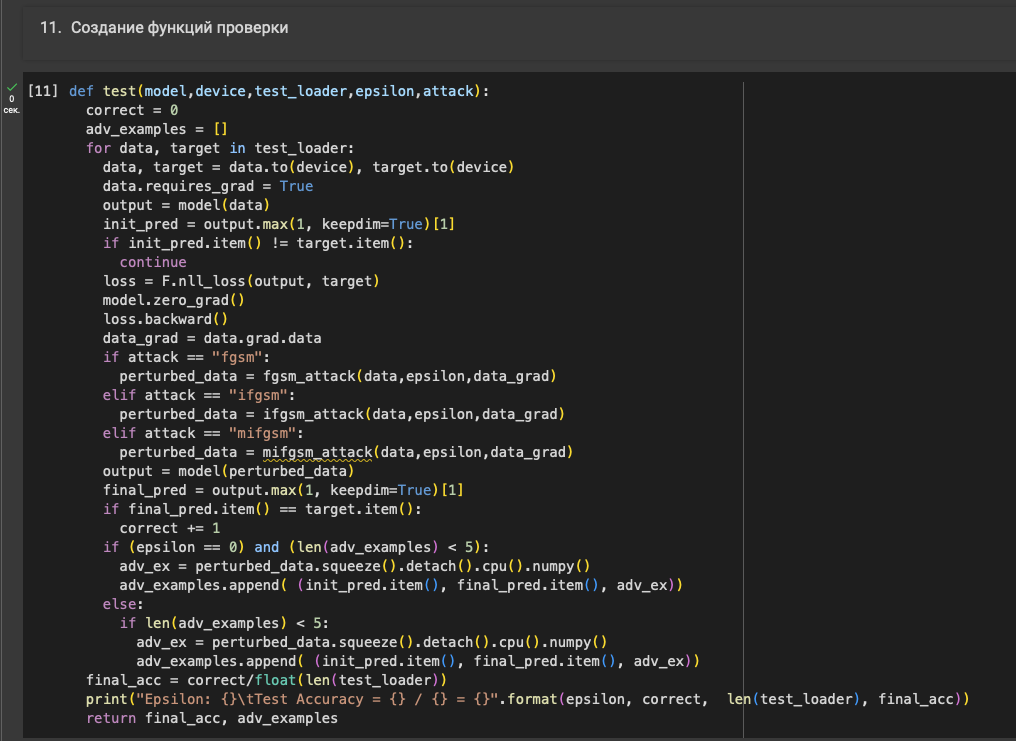
1. Построение графиков потерь при обучении и валидации в зависимости от эпохи



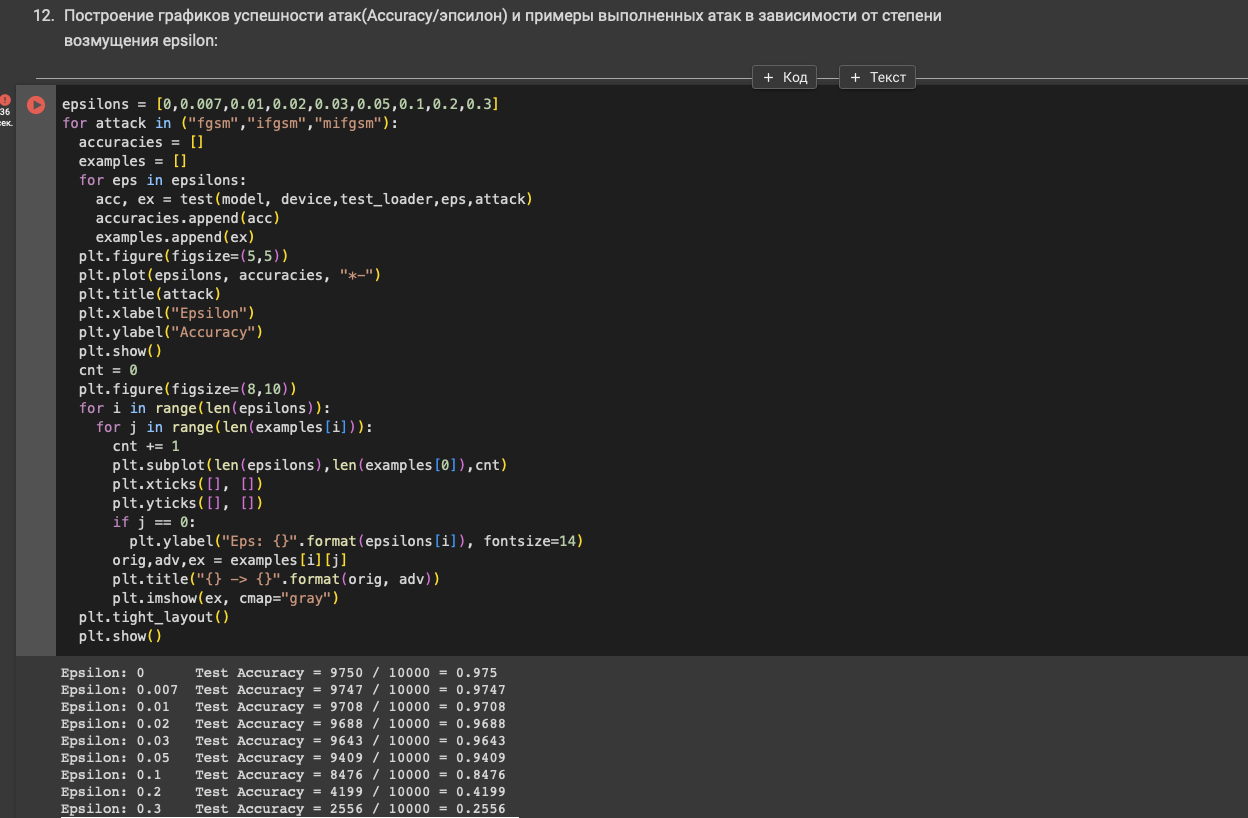
1. Создание функций атак FGSM, I-FGSM, MI-FGSM

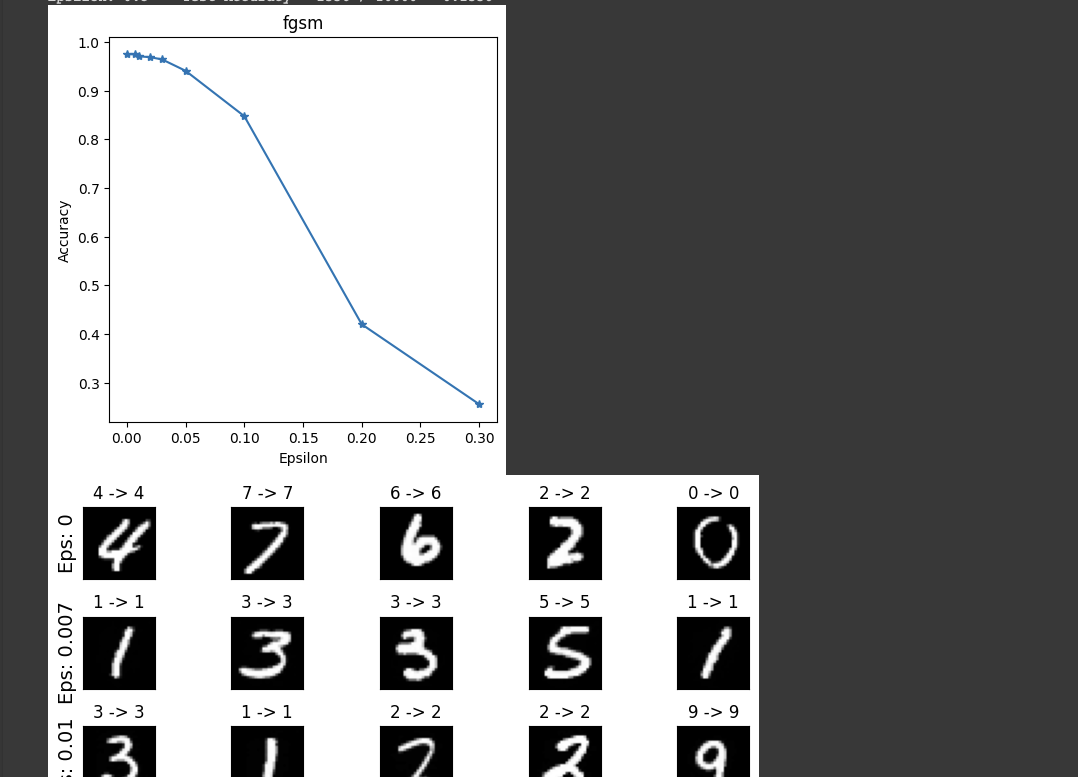


1. Создание функций проверки



1. Построение графиков успешности атак(Accuracy/эпсилон) и примеры выполненных атак в зависимости от степени возмущения epsilon(на моменте остановил расчет, так как гугл начал меня выбивать по OOM-киллеру. Остальная часть в jupyter-ноутбуке):

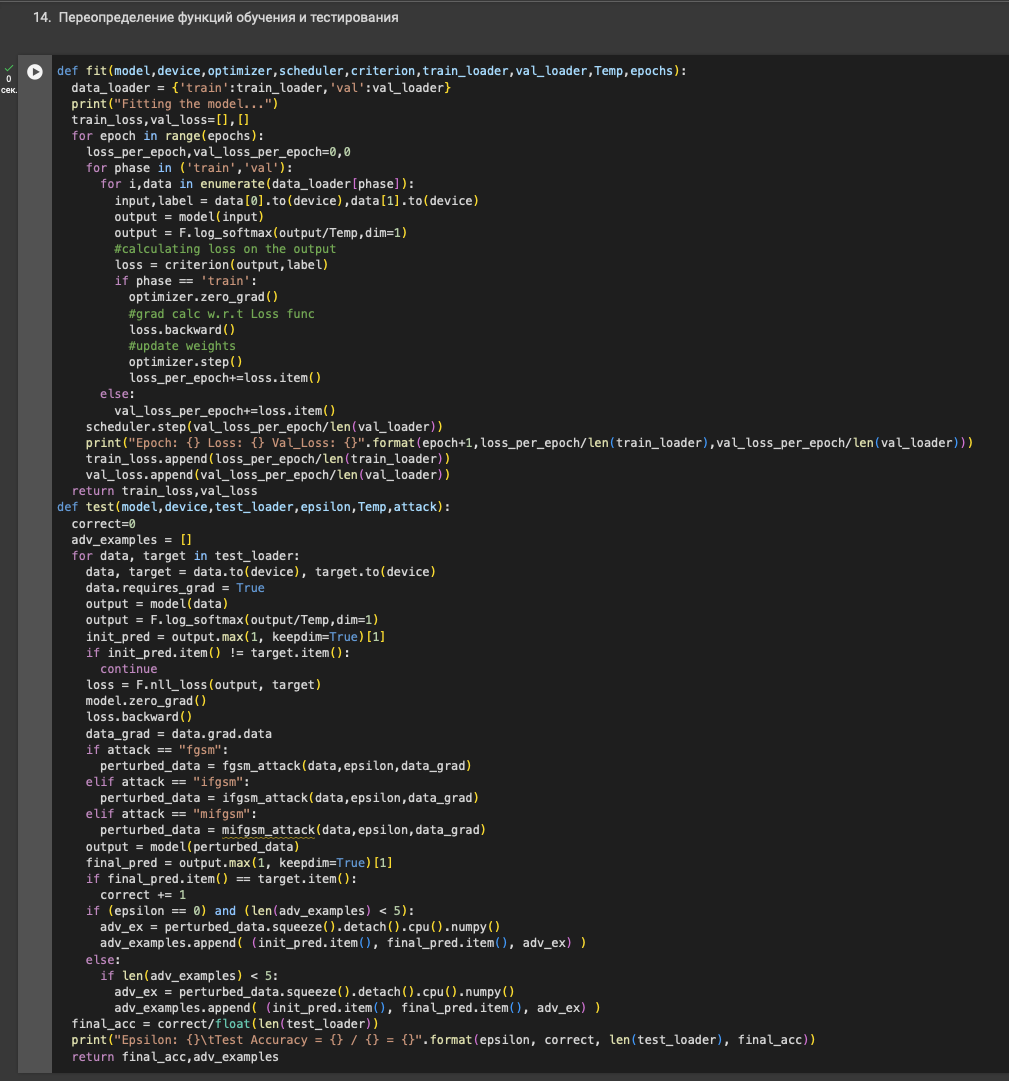




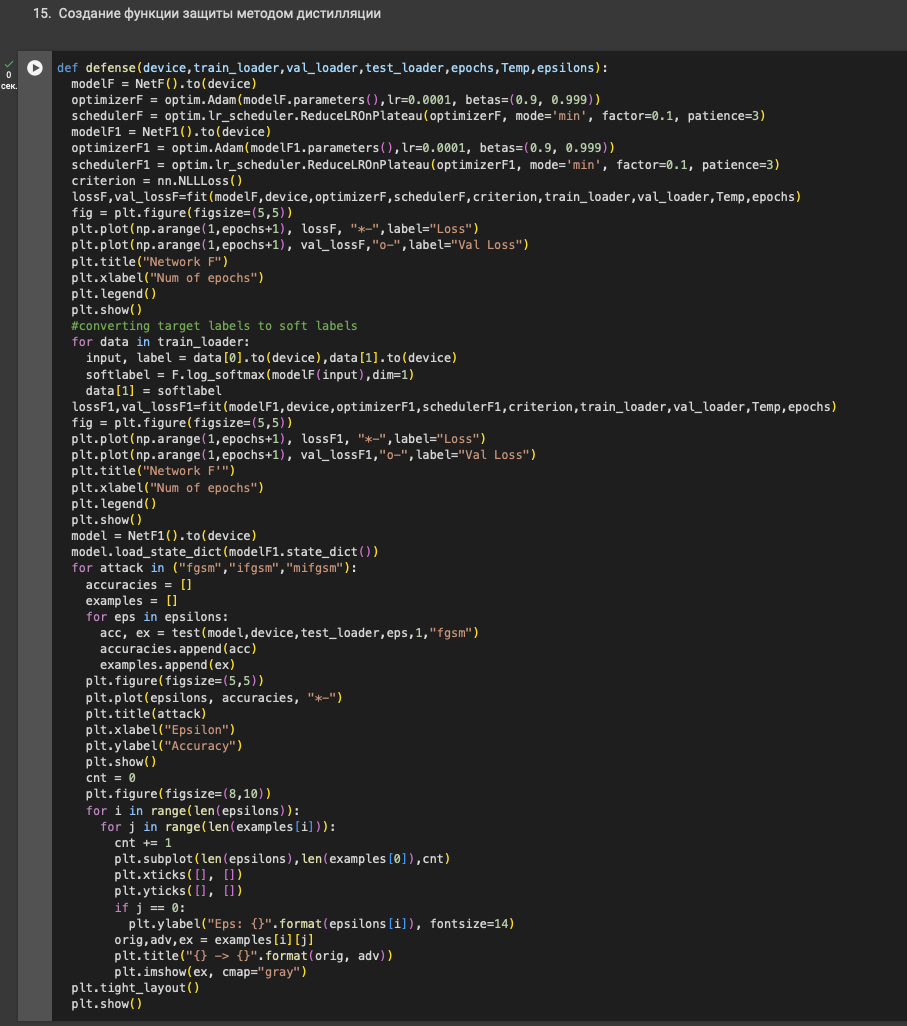
1. Создание двух классов НС



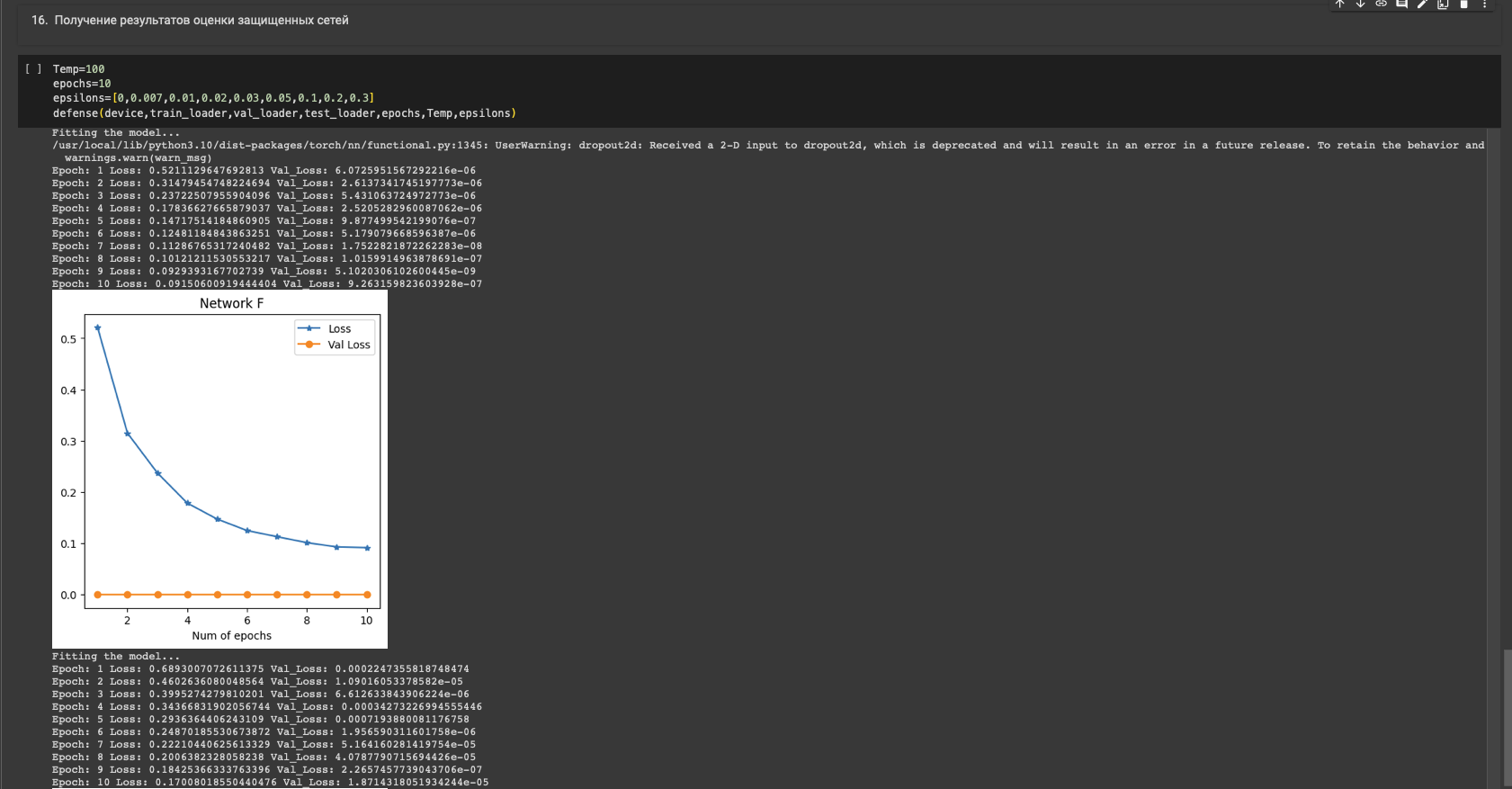
1. Переопределение функций обучения и тестирования



1. Создание функции защиты методом дистилляции



1. Получение результатов оценки защищенных сетей(остальная часть в jupyter-notebook)



**Вывод**

Метод защитной дистиляции (Protective Distillation) представляет собой стратегию обеспечения безопасности данных и информации от несанкционированного доступа или утечек. Этот метод основан на принципе разделения информации на несколько компонентов или элементов, где каждый из них не предоставляет полной или достаточно информативной картины без наличия остальных частей.

Основная идея защитной дистиляции заключается в том, что доступ к информации возможен только при наличии всех ее компонентов или элементов. Таким образом, информация остается безопасной даже в случае, если одна или несколько ее частей попадают в руки неавторизованного лица или подвергаются компрометации. Таким образом, данные становятся полезными только при условии, что все их элементы находятся в безопасности и доступны только авторизованным пользователям.

Оценка стойкости модели выглядит следующим образом:

- Атака fgsm снизила точность обычной модели до -21%, защищенной - до -90%;

- Атака ifgsm снизила точность не защищенных данных до -29%, защищенных - до -90%;

- Атака mifgsm снизила точность не защищенных данных до -29%, защищенных - до -91%.