|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования  **«МИРЭА – Российскийтехнологическийуниверситет»**  **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 – лабораторная работа №4

по дисципилне

«Анализ зазищенности систем искусственного интеллекта»

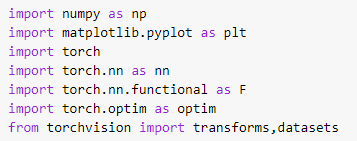
Группа: ББМО-01-22

Выполнила: Челышев Д.И.

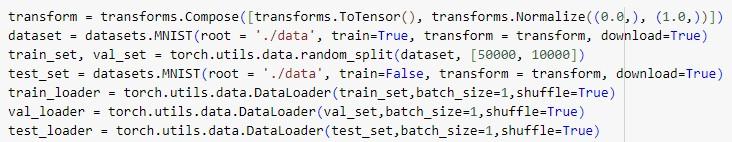
Проверил: Спирин А.А.

Москва 2023

Выполним импорт необходимых библиотек.



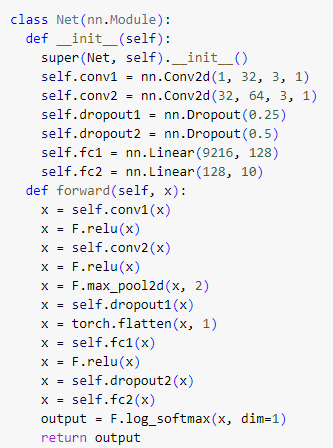
Зададим нормализующие преобразования, загрузим набор данных (MNIST), разобьем данные на подвыборки.



Настроим использование графического ускорителя.



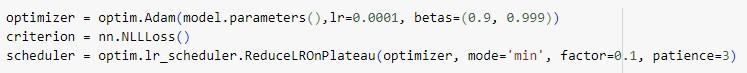
Создадим класс НС на основе фреймворка torch.



Проверим работоспособность созданного класса НС.



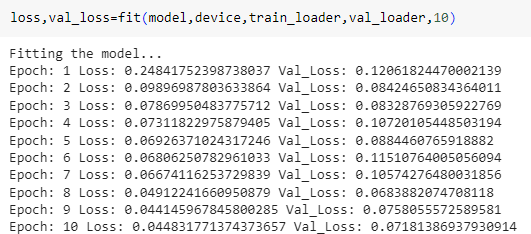
Создадим оптимизатор, функцию потерь и трейнер сети.



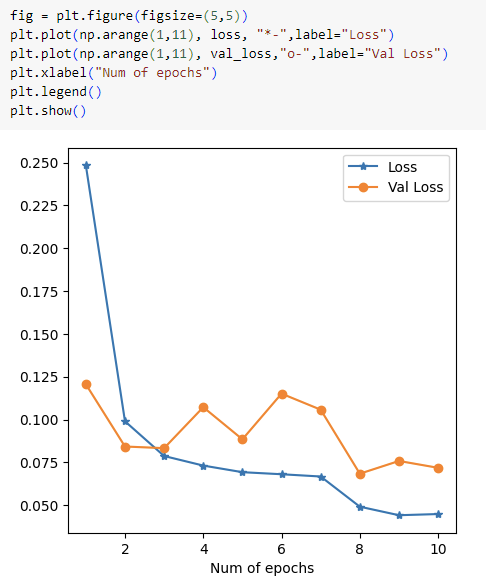
Определили функцию обучения сети.



Обучим модель.



Построим графики потерь при обучении и валидации в зависимости от эпохи.



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

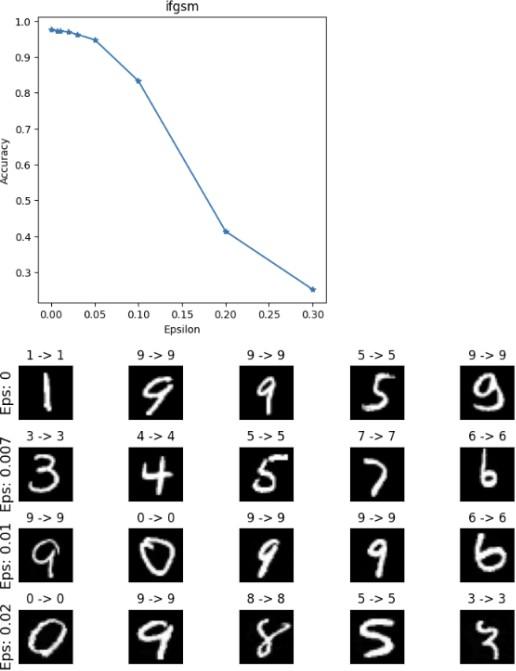
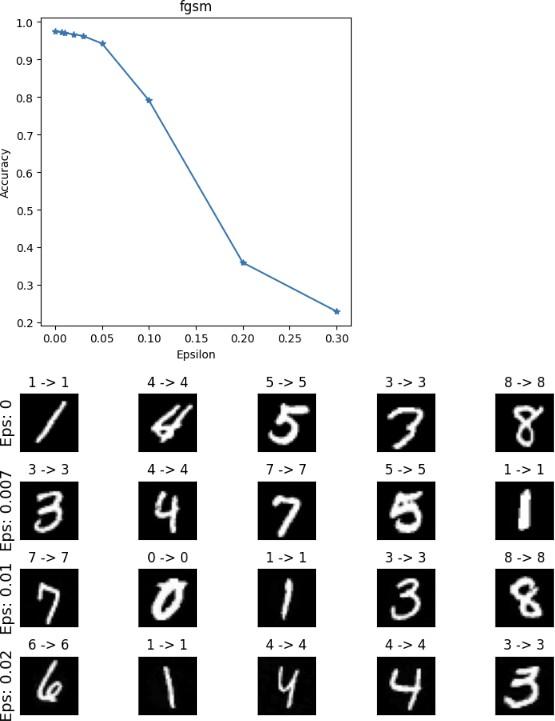


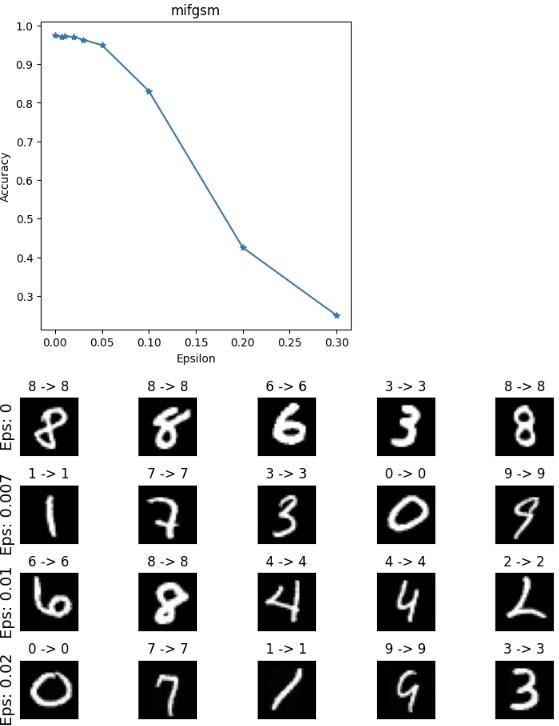
Создадим функцию проверки.



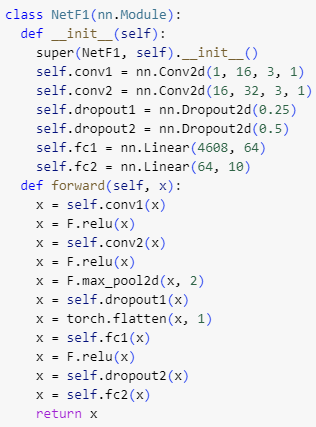
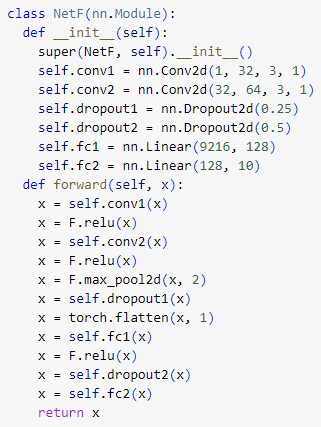
Построим графики успешности атак(Accuracy/эпсилон) и примеры выполненных атак в зависимости от степени возмущения epsilon.







Создадим 2 класса НС.



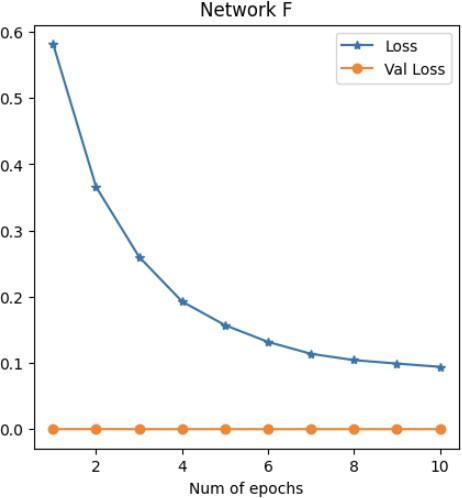
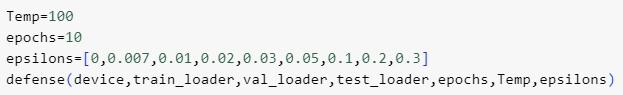
Переопределим функцию обучения и тестирования.

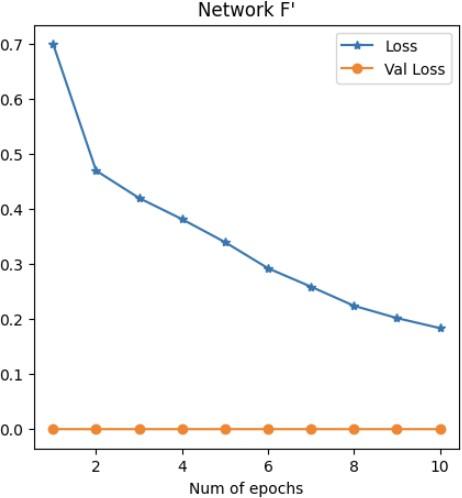


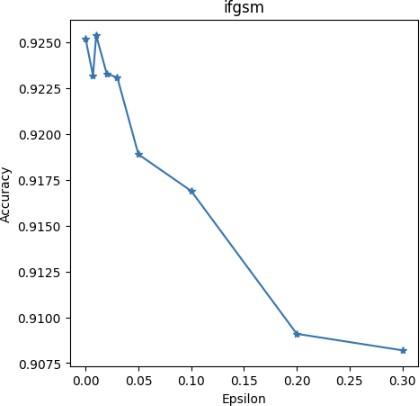
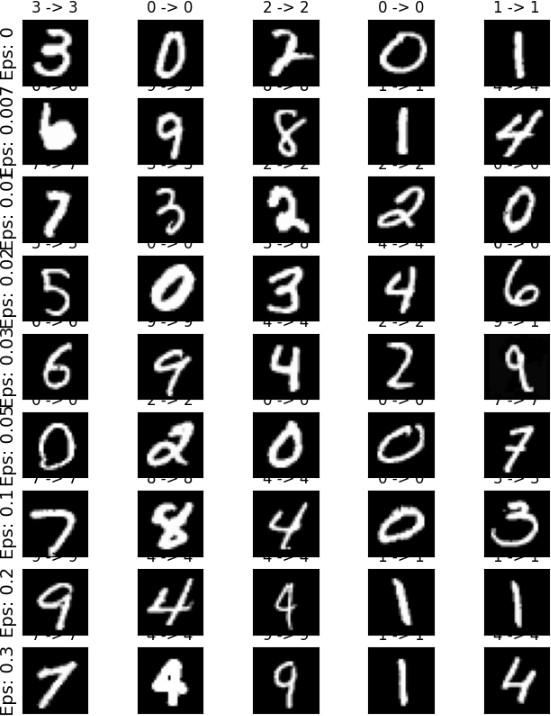
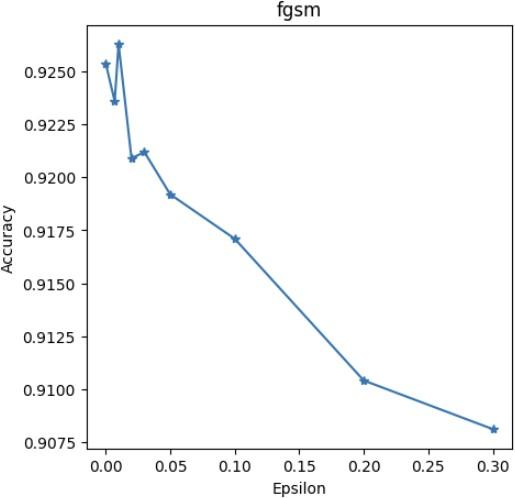
Создадим функцию защиты методом дистилляции.

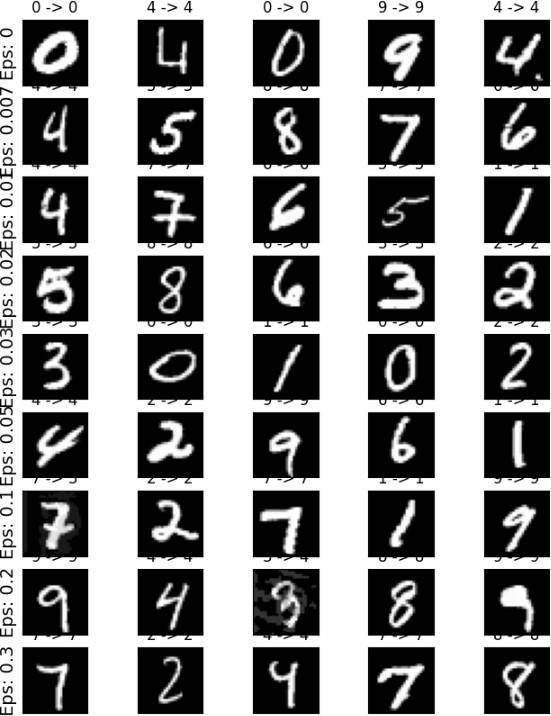


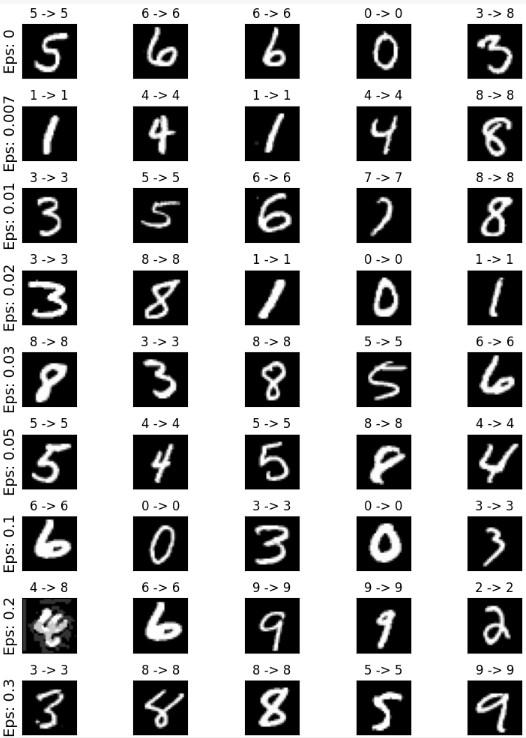
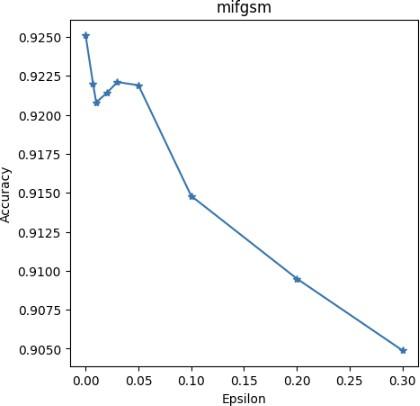
Получим результаты оценки защищенных сетей.











Вывод: Основная идея защитной дистиляции заключается в обучении устойчивой модели, путем передачи знаний от базовой модели,

подверженной атакам, к новой модели, которая спроектирована для устойчивости к различным атакам.

Дистилляция дает более плоские локальные минимумы. Следовательно, небольшие изменения во входных данных с меньшей вероятностью изменят прогнозируемые значения.

Итог по увеличению стойкости модели:

атака fgsm снизила точность не защищенных данных до - 14%, защищенных - до - 91%;

атака ifgsm снизила точность не защищенных данных до - 15%, защищенных - до - 91%;

атака mifgsm снизила точность не защищенных данных до - 15%, защищенных - до - 91%.

Модель, которая обучалась на метках учителя имеет большее значение потерь после обучения, но это невилируется стойкостью к атакам FGSM.