|  |  |  |
| --- | --- | --- |
|  |  |  |
| 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 №2**

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

Выполнил:

Евдокимов А.М.

Группа: ББМО-02-23

Москва - 2024

**Задачи:** 1. Реализовать атаки уклонения на основе белого ящика против классификационных моделей на основе глубокого обучения.

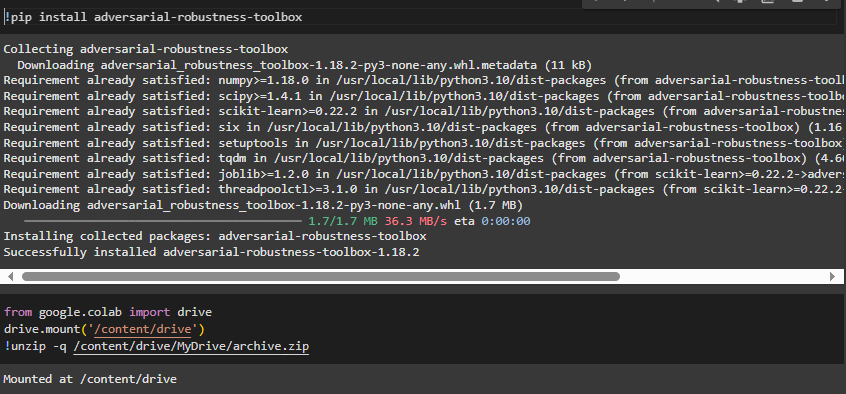
2. Получить практические навыки переноса атак уклонения на основе черного ящика против моделей машинного обучения.

3. Реализовать атаки уклонения на основе белого ящика против классификационных моделей на основе глубокого обучения.

**Набор данных:** Для этой части используйте набор данных GTSRB (German Traffic Sign Recognition Benchmark). Набор данных состоит примерно из 51 000 изображений дорожных знаков. Существует 43 класса дорожных знаков, а размер изображений составляет 32×32 пикселя.

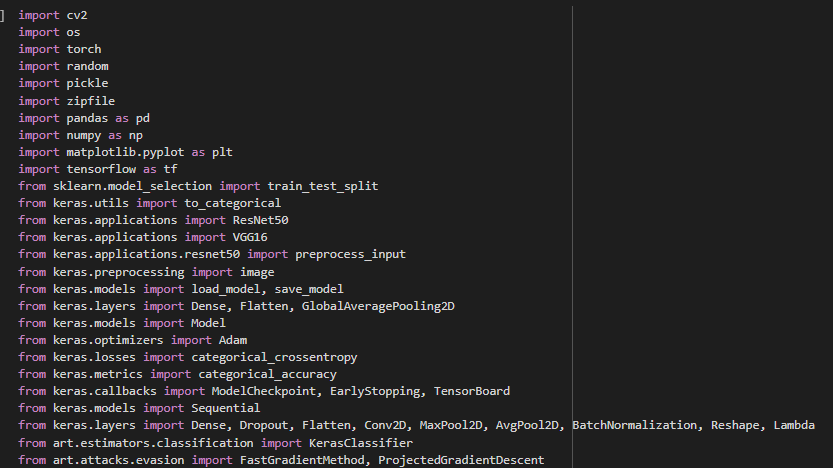
**Выполнение лабораторной работы**

**Задание 1. Обучить 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB**

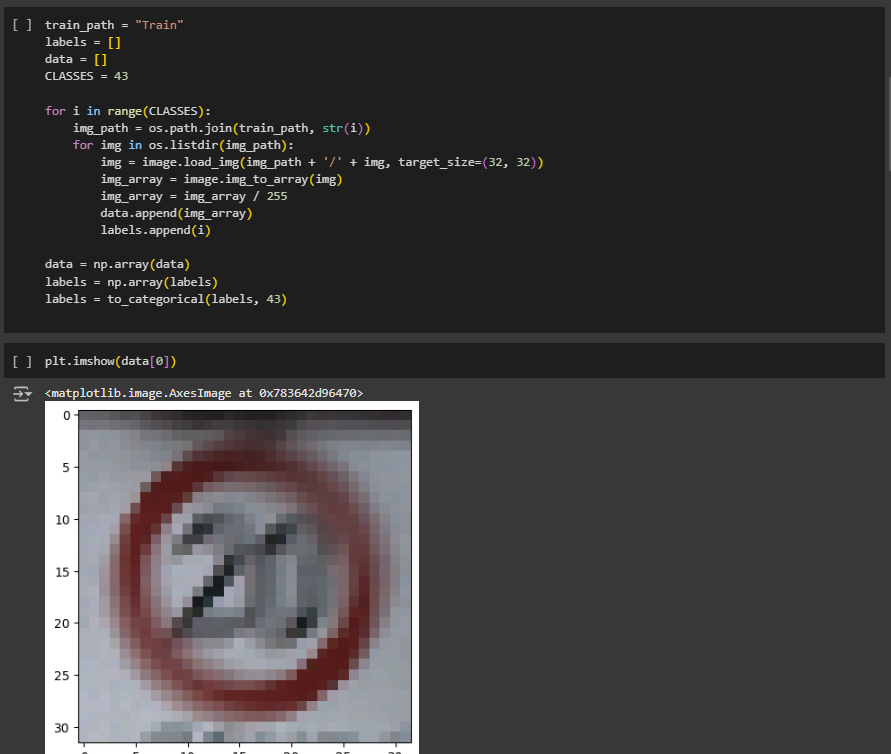


Скачаем набор данных c дорожными знаками по ссылке https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german traffic-sign/ и загрузим в среду Google Colab.

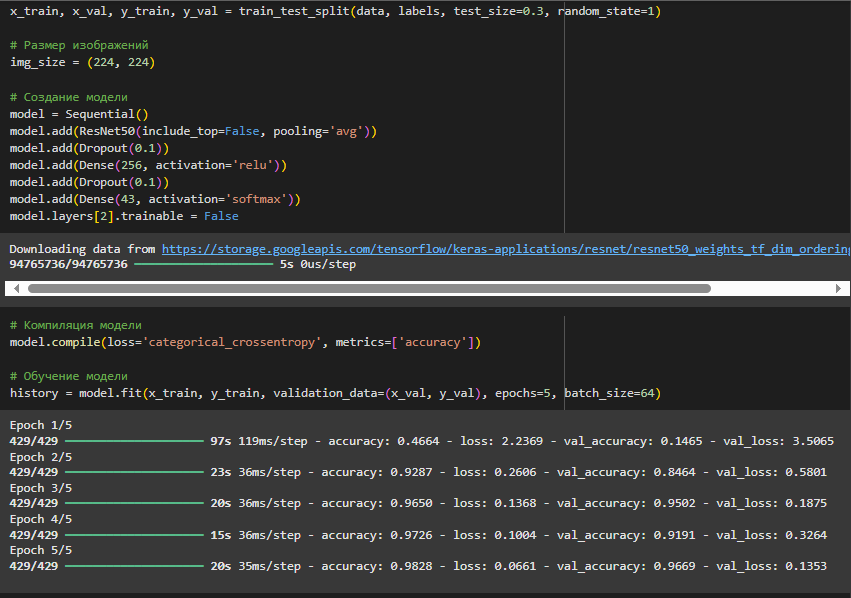
Выполним импорт необходимых библиотек:



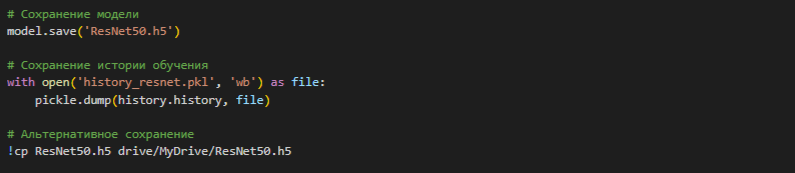
Извлечём изображения для создания тренировочной выборки:



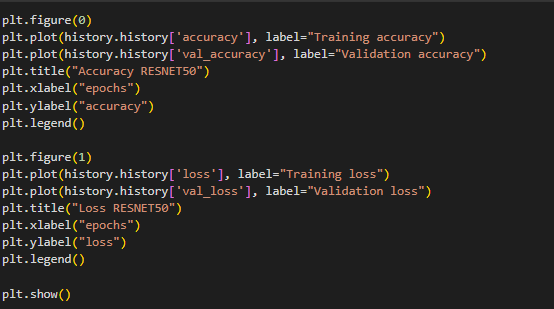
Воспользуемся ResNet50. Разобьём датасет на тренировочную и тестовую выборки в соотношении 70:30 и поменяем выходные слои модели, для осуществления классификации 43 типов изображений:

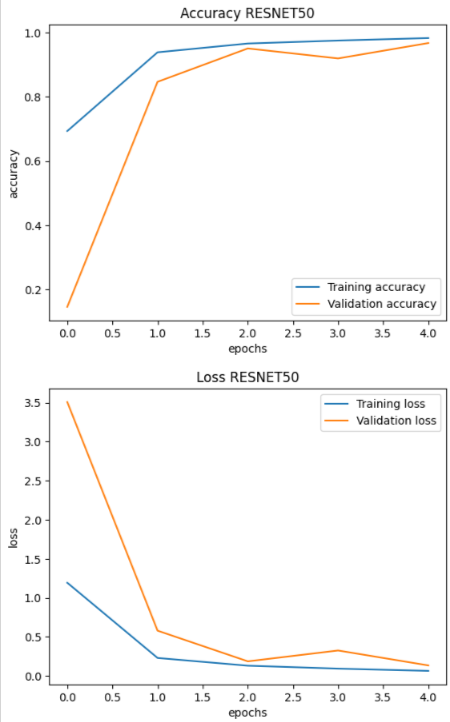
****

Обучим изменённую модель с параметрами epochs = 5, batch\_size = 64 и сохраним модель.

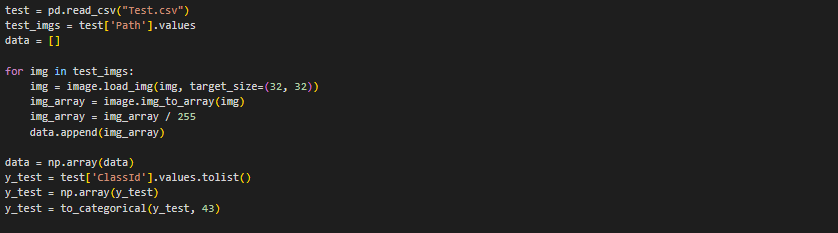


Построим два графика, которые отражают успешность обучения модели ResNet50 с изменёнными выходными слоями:

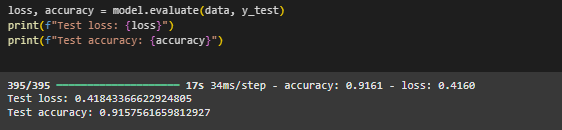




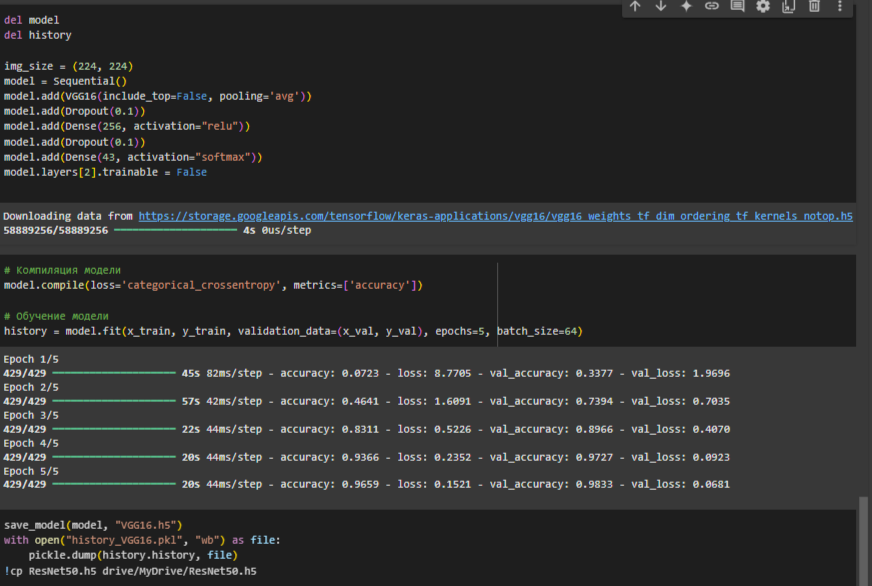
Скорректируем тестовый набор данных



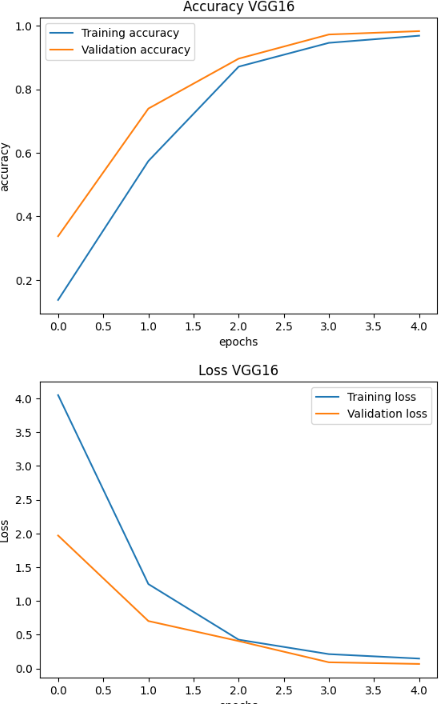
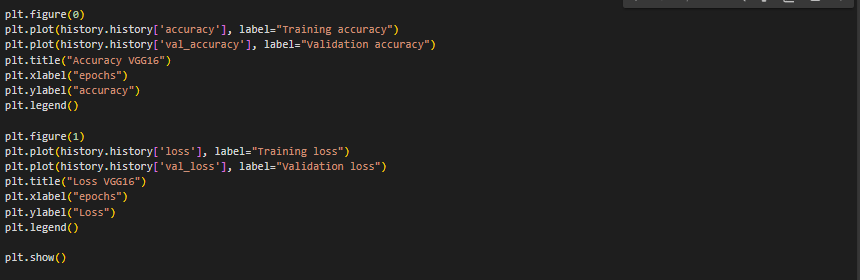
Оценим точность классификации модели:



Выполним тоже самое для VGG16:



Визуализация

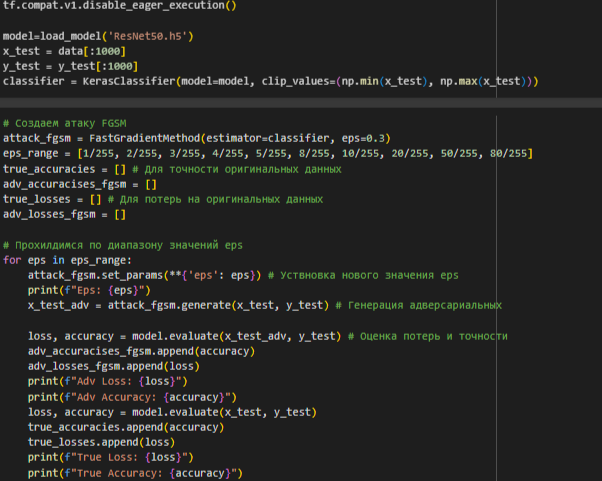


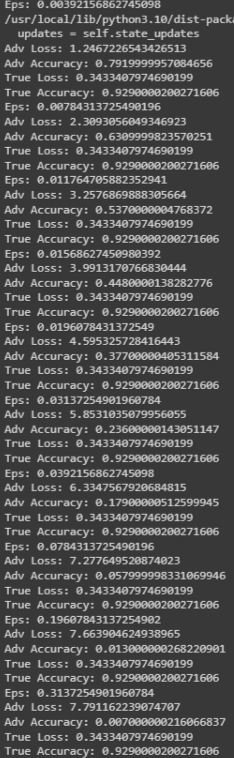
Занесём результаты обучений, валидаций и тестов в сравнительную таблицу 1.

Таблица 1 – Сравнительная таблица

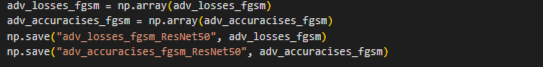
|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Обучение | Валидация | Тест |
| ResNet50 | accuracy: 0.9828 loss: 0.0661 | val\_accuracy: 0.9669 val\_loss: 0.1353 | Test accuracy: 0.9158  Test loss: 0.4184 |
| VGG16 | accuracy: 0.9659 loss: 0.1521 | val\_accuracy: 0.9833 val\_loss: 0.0681 | Test accuracy: 0.9492  Test loss: 0.2575 |

**Задание 2. Применение нецелевой атаки уклонения на основе белого ящика против моделей глубокого обучения**

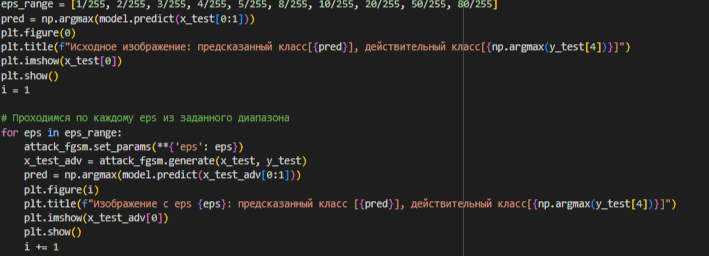
****

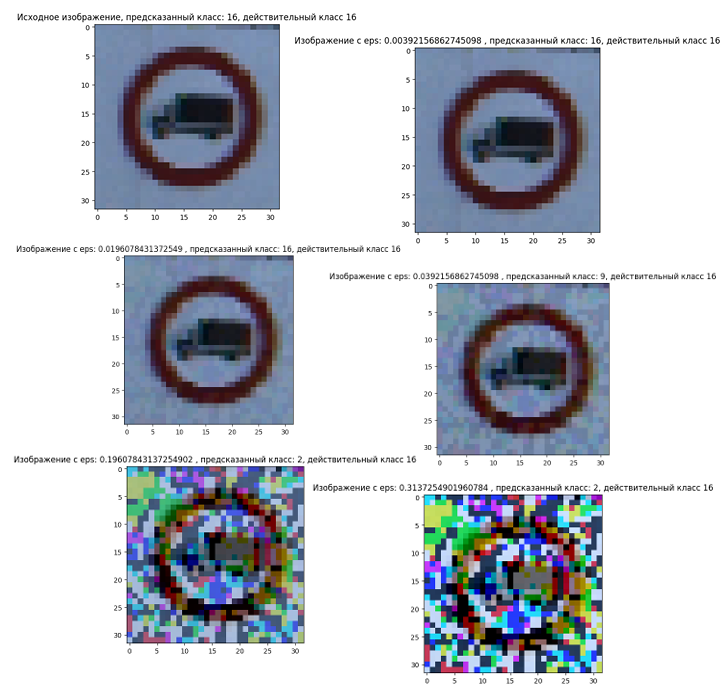


Сохраним атаку FGSM для дальнейшего анализа

****

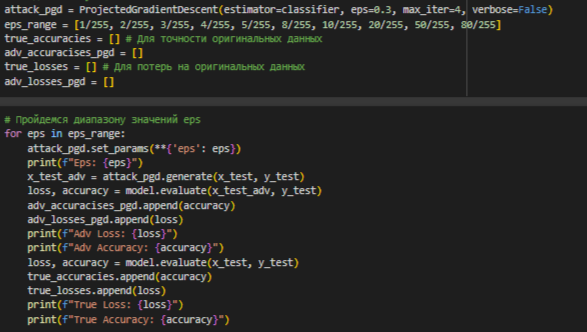
Отображаем исходные и адверсариальные изображения для разных значений eps

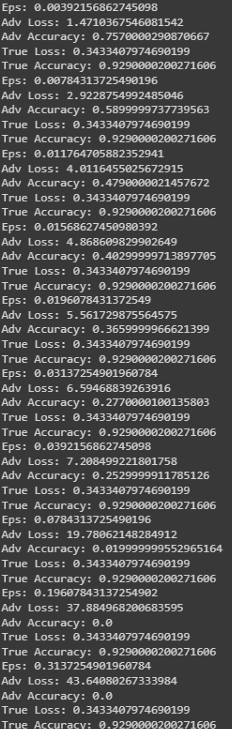
****

****

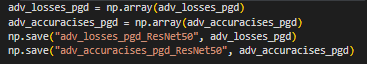
Видно, что при росте eps, шум на картинке сильно увеличивается, и с 5/255 уже становится более заметен. Оптимальным eps будет значение от 5/255 до 10/255.

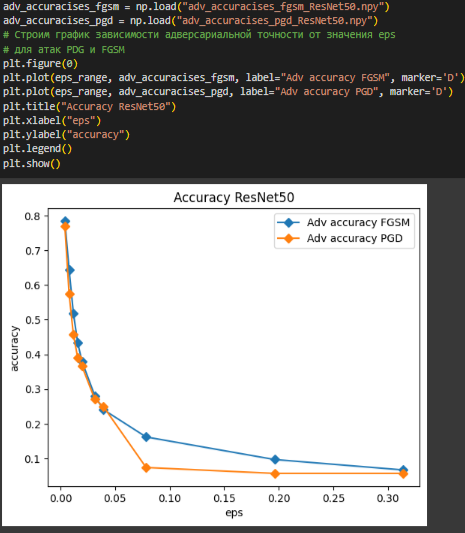
Теперь реализуем атаку PGD на ResNet50:

****

****

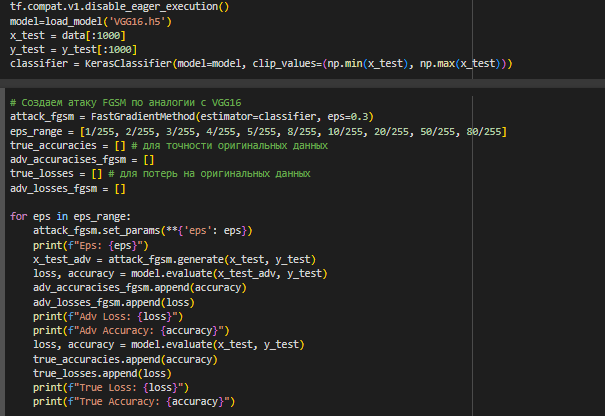
Сохраним атаку PGD для дальнейшего анализа с помощью графика

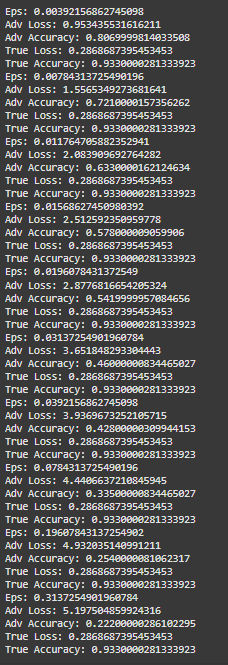
****

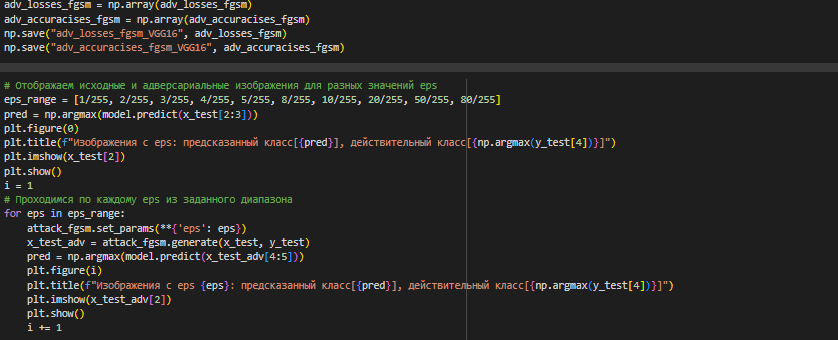
****

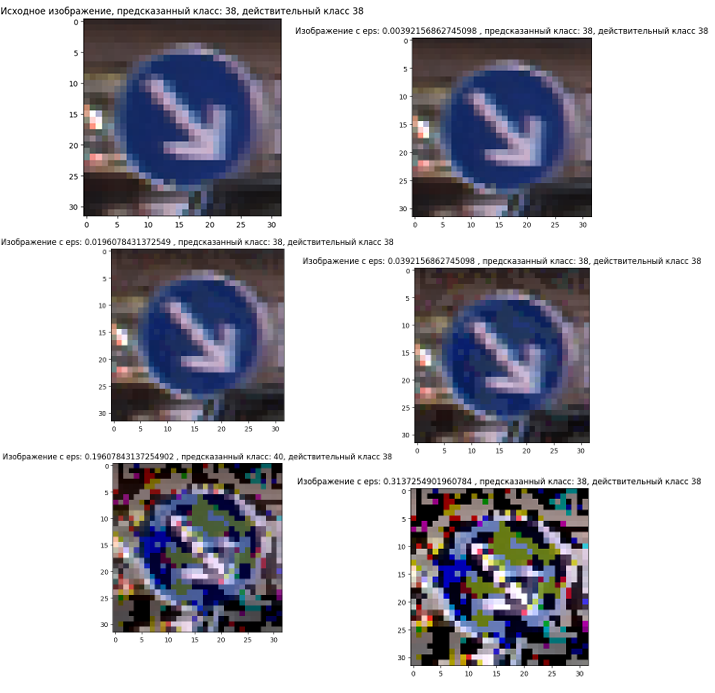
Из графиков видно, что методы имеют почти схожую эффективность, но метод PGD слегка больше снижает точность.

Реализуем атаку FGSM на VGG16:

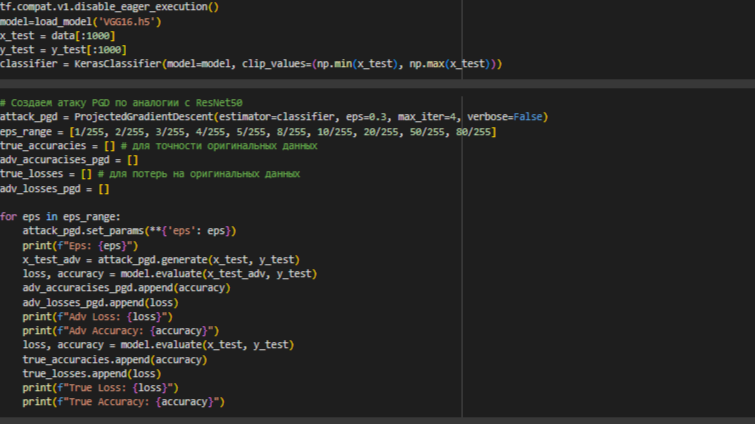
****

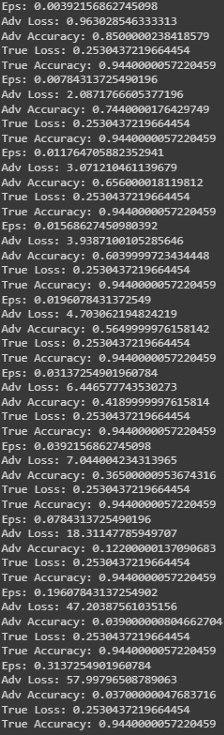
****

****

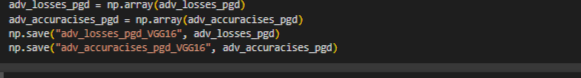
****

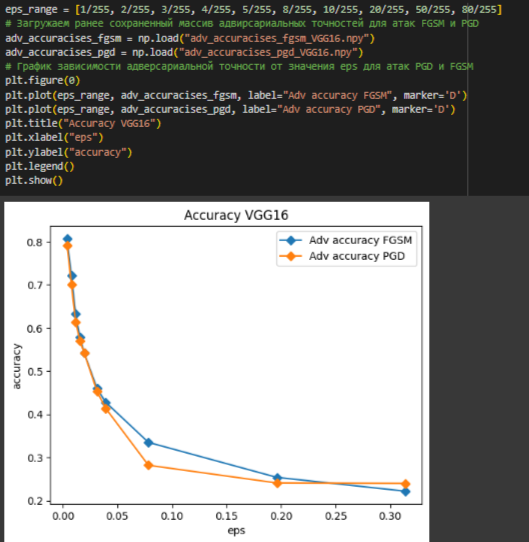
Выполним атаку PGD на VGG16:





Сохраним атаку PGD для дальнейшего анализа с помощью графика





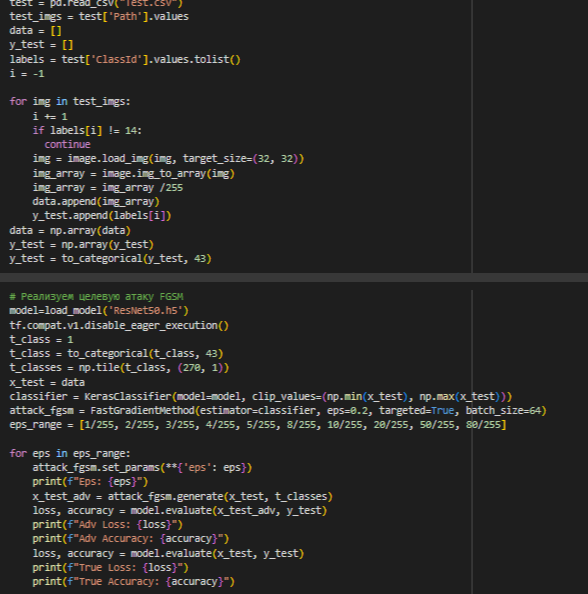
Из графиков видно, что методы имеют почти схожую эффективность, но метод PGD слегка больше снижает точность.

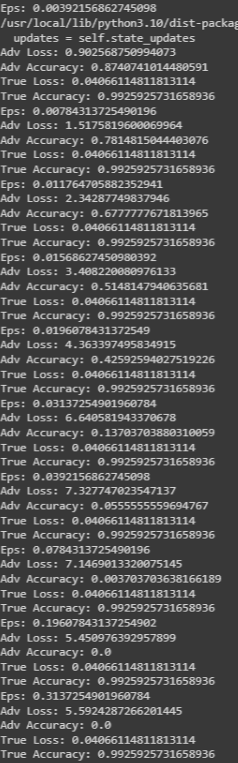
Таблица 2 - Зависимость точности классификации от параметра искажений eps

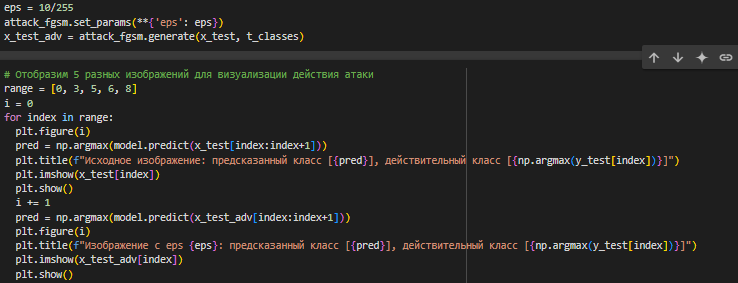
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Исходные изображения | Adversarial images 𝜖=1/255 | Adversarial images 𝜖=1/255 | Adversarial images 𝜖=1/255 |
| ResNet50 - FGSM | 0,9221 | 0,7920 | 0,3370 | 0,1790 |
| ResNet50 - PGD | 0,9221 | 0,7570 | 0,3360 | 0,2530 |
| VGG16 - FGSM | 0,9465 | 0,8650 | 0,4930 | 0,2530 |
| VGG16 - PGD | 0,9465 | 0,8500 | 0,5650 | 0,3650 |

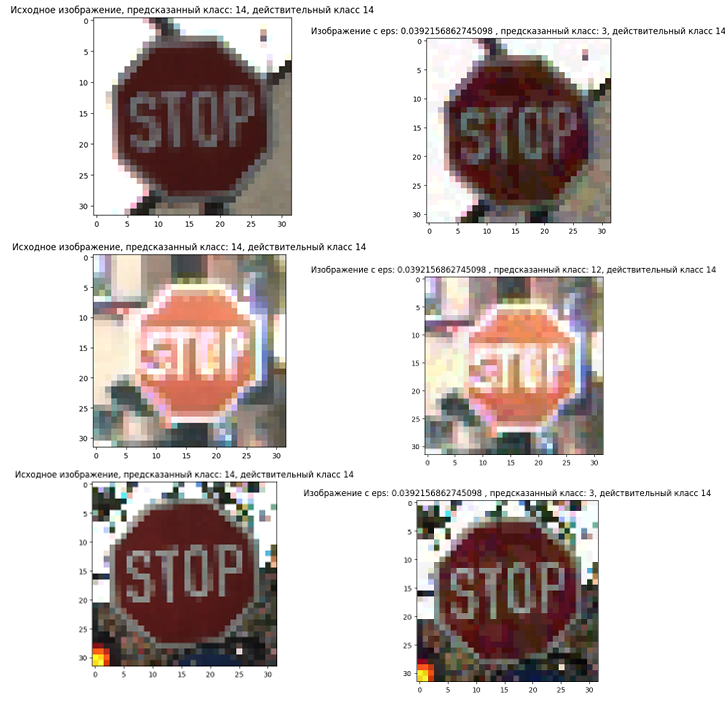
**Задание 3. Применение целевой атаки уклонения методом белого ящика против моделей глубокого обучения**

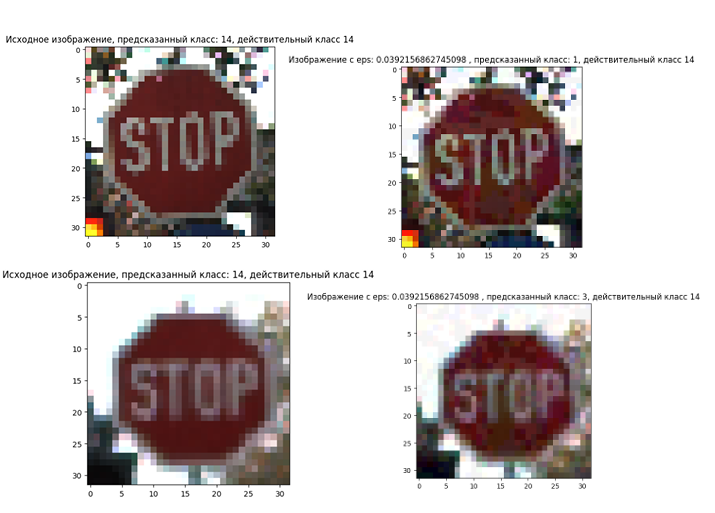
Выполним целевую атаку FGSM на ResNet50:



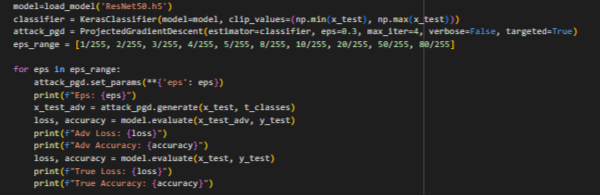
****

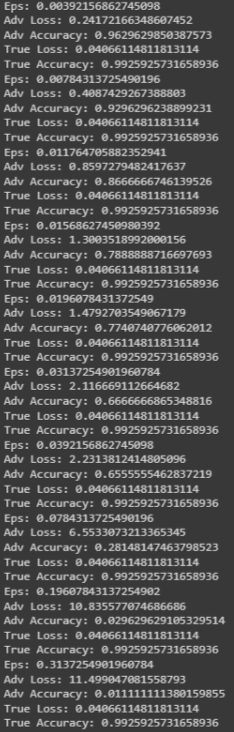
****

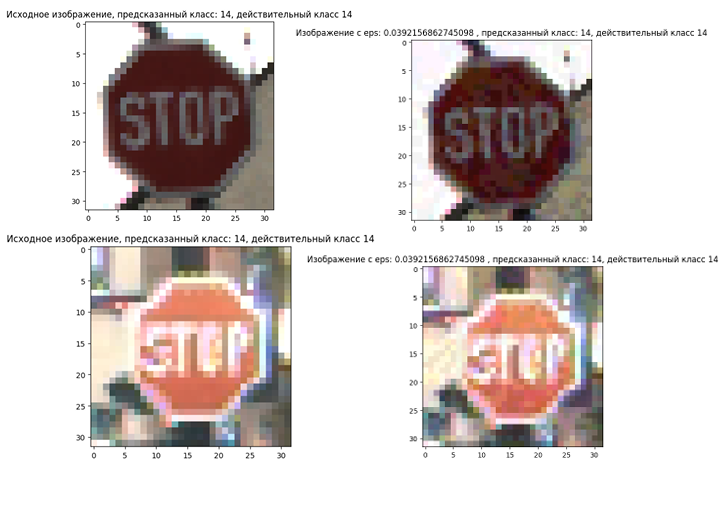
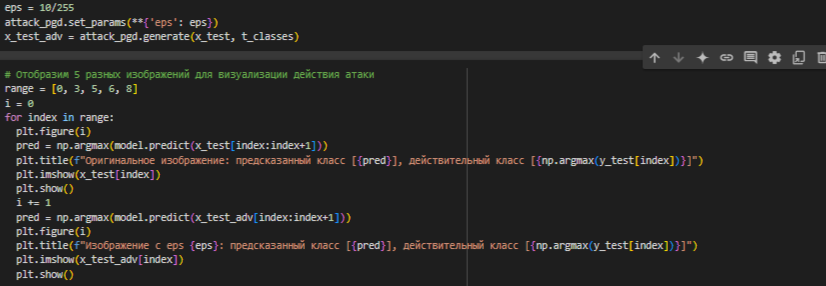
****

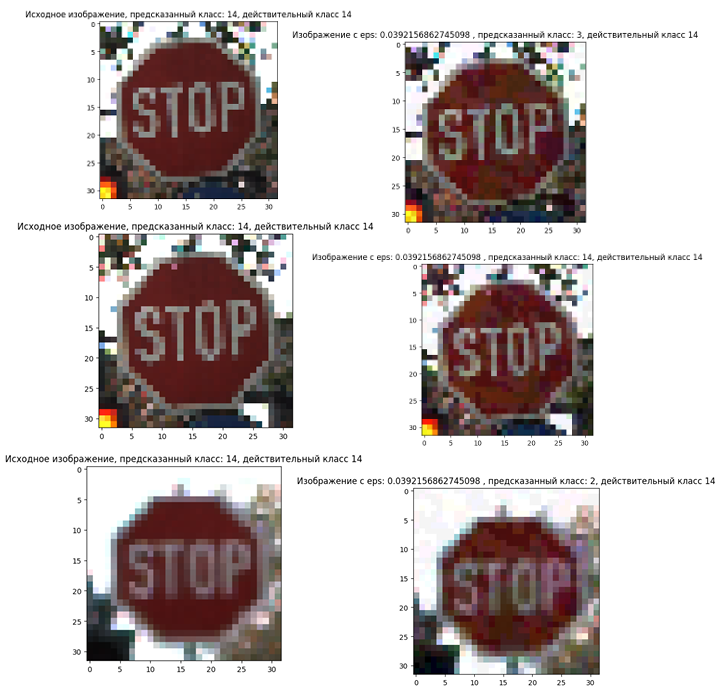
****

Выполним целевую атаку PGD на ResNet50:





****

****

Заполним таблицу 3, в которой представим точность целевых атак PGD и FGSM на знак стоп

Таблица 3 – Точность целевых атак

|  |  |  |
| --- | --- | --- |
| Искажение | PGD attack – Stop sign images | FGSM attack – Stop sign images |
| ε =1/255 | 0,9630 | 0,8741 |
| ε =1/255 | 0,8667 | 0,6778 |
| ε =1/255 | 0,7741 | 0,4259 |
| ε =1/255 | 0,6556 | 0,5556 |
| ε =1/255 | 0,2815 | 0,037 |
| ε =1/255 | 0,0296 | 0 |
| ε =1/255 | 0,0111 | 0 |

**Результаты:**

По результатам проведенных атак можем сделать вывод, что метод FGSM неэффективен при целевых атаках, поскольку при увеличении искажения возникают ошибки в классификации. Оптимальное значение искажения составляет 10/255; превышение этого значения приводит к постоянным ошибкам модели.

В отличие от этого, метод PGD прекрасно подходит для целевых атак. Даже при значительных искажениях модель чаще всего правильно определяет заданный класс, но изображение становится слишком явно искаженным. Оптимальное значение искажения составляет 10/255.

**Заключение**

В ходе выполнения лабораторной работы были достигнуты следующие результаты. Были разработаны и обучены два классификатора на основе глубоких нейронных сетей (ResNet50 и VGG16) на датасете GTSRB. Проведён анализ их производительности, который показал высокую точность моделей: для ResNet50 точность тестирования составила 91,58%, а для VGG16 - 94,92%. Реализованы нецелевые атаки уклонения FGSM и PGD для моделей ResNet50 и VGG16. Построены графики зависимости точности классификации от параметра искажения, которые показали, что обе атаки эффективно снижают точность моделей, при этом PGD оказался более эффективным при увеличении параметра искажения. Также проведены целевые атаки FGSM и PGD на изображениях дорожного знака "Стоп" для их классификации как знака "Ограничение скорости 30".

В процессе экспериментов метод FGSM продемонстрировал низкую устойчивость при значительных искажениях, тогда как PGD показал лучшую производительность, хотя и с сильными визуальными искажениями изображений. Полученные результаты подтвердили уязвимость моделей глубокого обучения к атакам уклонения, что подчёркивает необходимость разработки механизмов защиты, таких как более устойчивые архитектуры, регуляризация или алгоритмы обнаружения атак. Лабораторная работа позволила глубже изучить уязвимости современных моделей и получить практические навыки реализации атак уклонения.