

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

# «МИРЭА – Российский технологический университет»

РТУ МИРЭА

ИКБ направление «Киберразведка и противодействие угрозам с применением технологий искусственного интеллекта» 10.04.01

Кафедра КБ-4 «Интеллектуальные системы информационной безопасности»

Лабораторная работа №2

по дисциплине

Анализ защищенности систем искусственного интеллекта»

Группа: ББМО-02-22

Выполнил:

Поплавский

М.С.

Проверил: к.т.н Спирин А.А.

Москва 2023

# Задачи

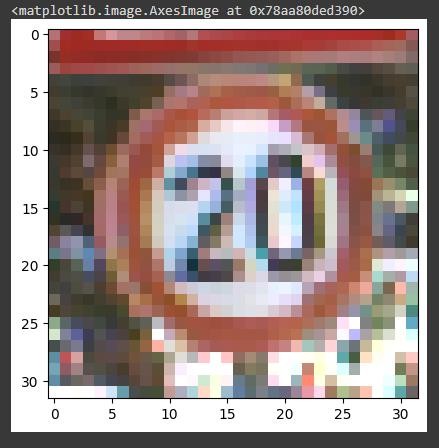
* Реализовать атаки уклонения на основе белого ящика против классификационных моделей на основе глубокого обучения.
* Получить практические навыки переноса атак уклонения на основе черного ящика против моделей машинного обучения.

# Набор данных

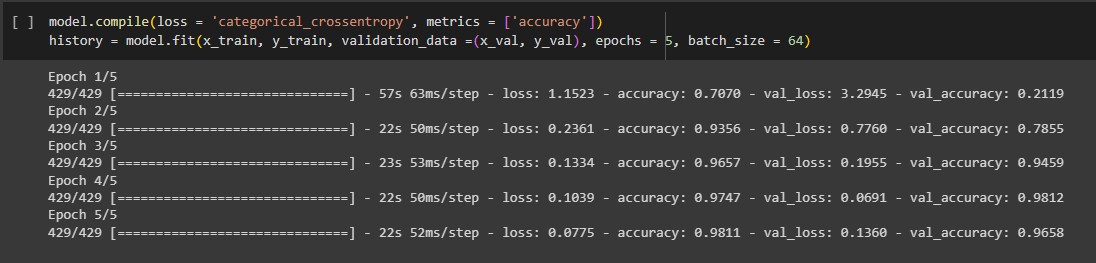
Для этой части используйте набор данных GTSRB (German Traffic Sign Recognition Benchmark). Набор данных состоит примерно из 51 000 изображений дорожных знаков. Существует 43 класса дорожных знаков, а размер изображений составляет 32×32 пикселя.

# Задание 1

Обучим 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB. При извлечении картинок для создания тренировочной выборки, получим матричное представление картинки. Для восприятия моделями нейронных сетей, данные были масштабированы.

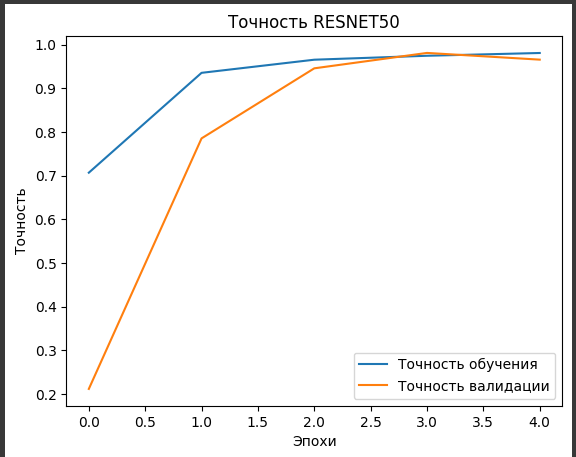


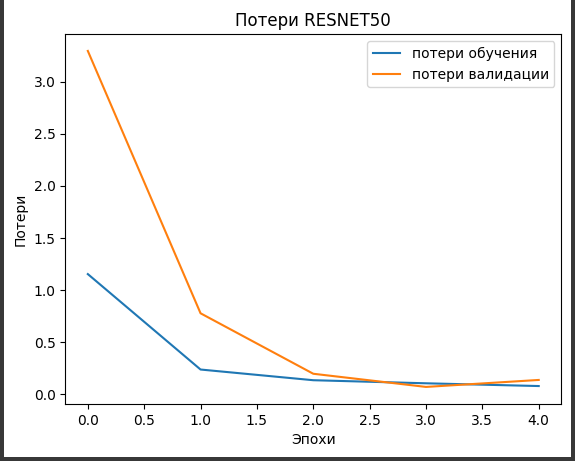
В качестве первой модель будет использоваться ResNet50. Опытным путём были выбраны лучшие значения количества эпох обучения и размера пакета.



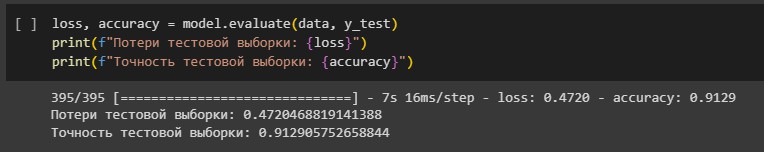
# RestNet50

Далее необходимо построить графики, отражающие качество обучения модели ResNet50. Было принято решение остановиться на 5 эпохах, так как итоговая точность увеличилась по мере роста числа эпох.





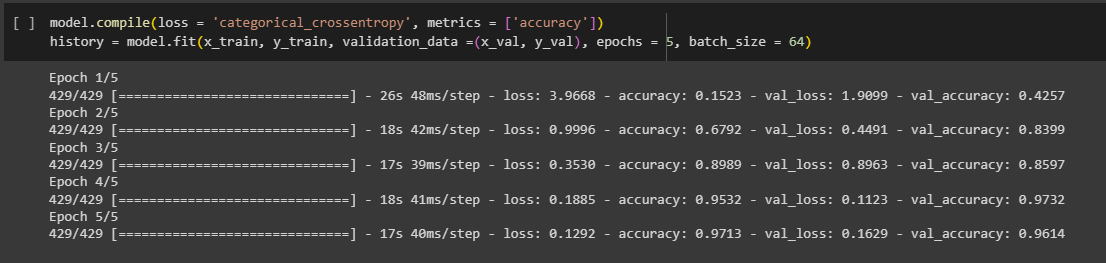
После этого необходимо проверить модель на тестовом наборе данных



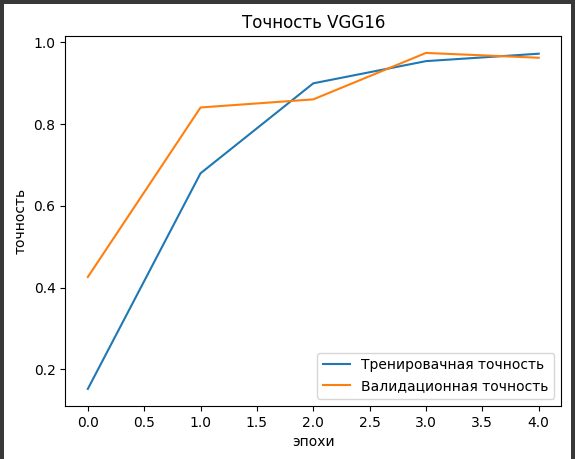
Точность составила 91%

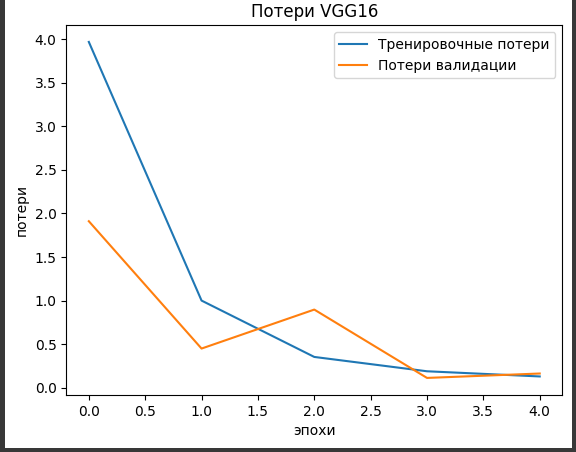
# VGG16

Следующей будет модель VGG16, необходимо её обучить

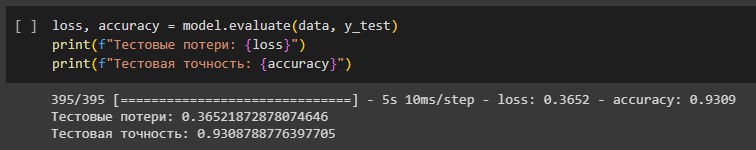


Далее необходимо построить графики, аналогичные графикам для модели ResNet50





После этого необходимо проверить модель на тестовом наборе данных



Точность составила 93%

Таблица 1. Сравнительная таблица моделей ResNet50 и VGG16

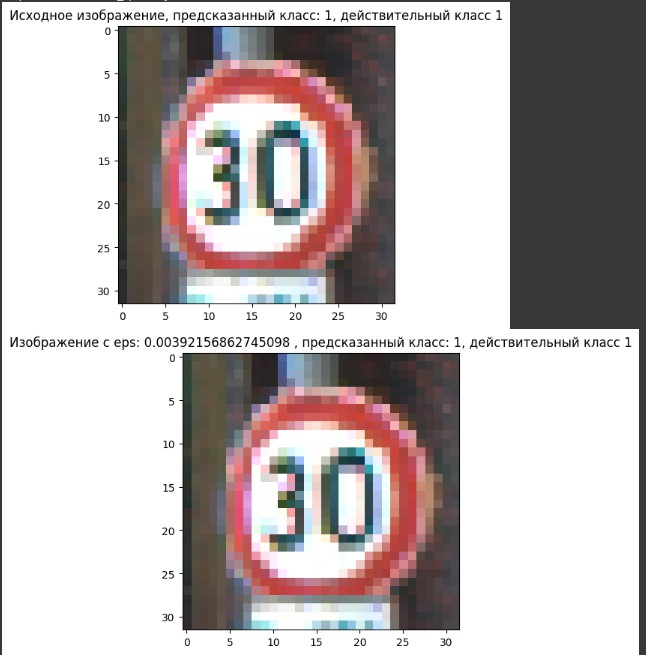
|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Обучение** | **Валидация** | **Тест** |
| ResNet50 | Потери: 0.08 | Потери: 0.14 | Потери: 0.47 |
| Точность: 0.98 | Точность: 0.97 | Точность: 0.91 |
| VGG16 | Потери: 0.13 | Потери: 0.16 | Потери: 0.37 |
| Точность: 0.97 | Точность: 0.96 | Точность: 0.93 |

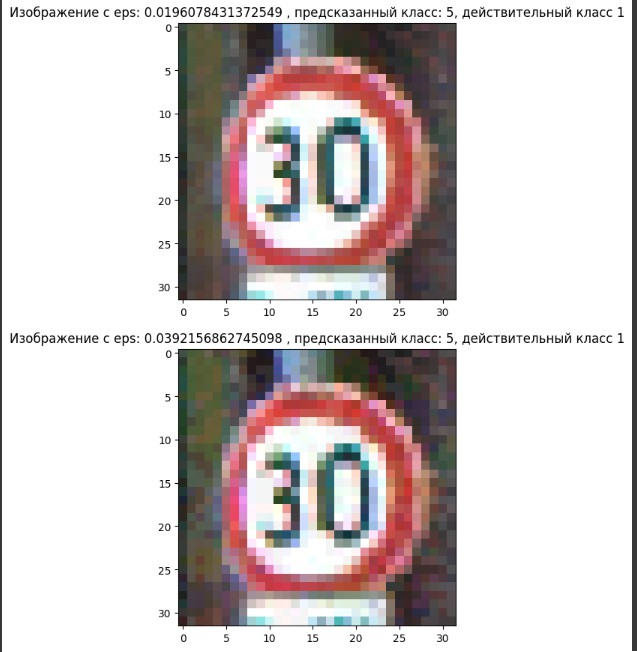
# Задание 2

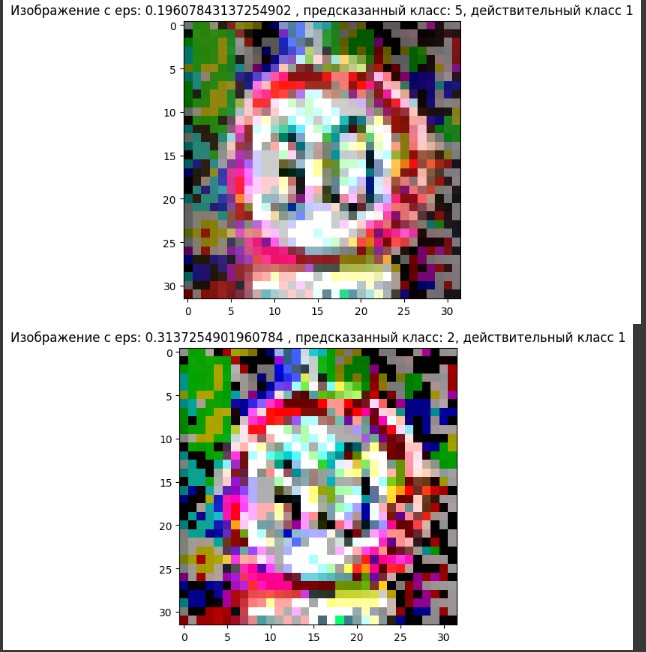
Применить нецелевую атаку уклонения на основе белого ящика против моделей глубокого обучения. Реализовать следующие типы атак:

* Fast Gradient Sign Method (FGSM)
* Projected Gradient Descent (PGD)

Необходимо создать модель атаки, которая основывается на классификаторе для внесения шума в изображение. Ниже представлено отображение исходного и атакующих изображений для атаки FGSM

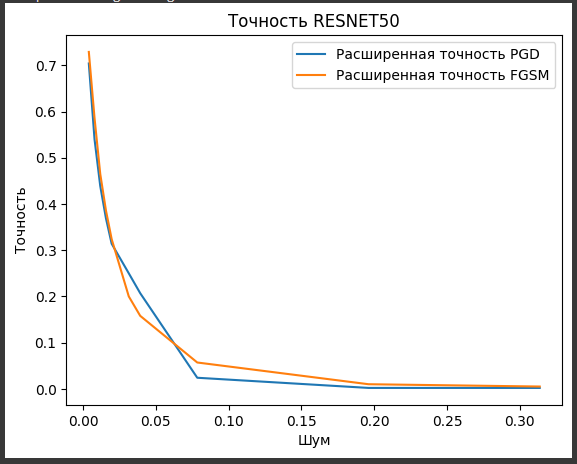




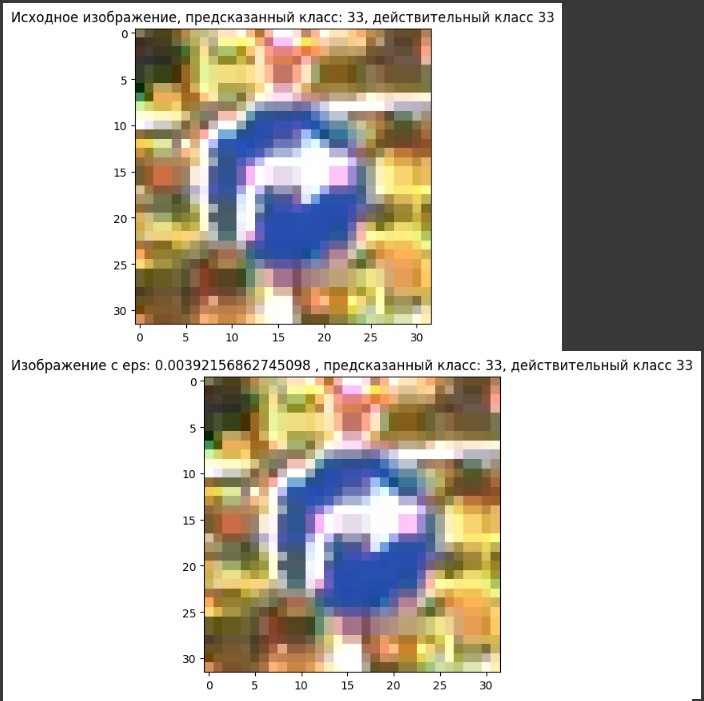


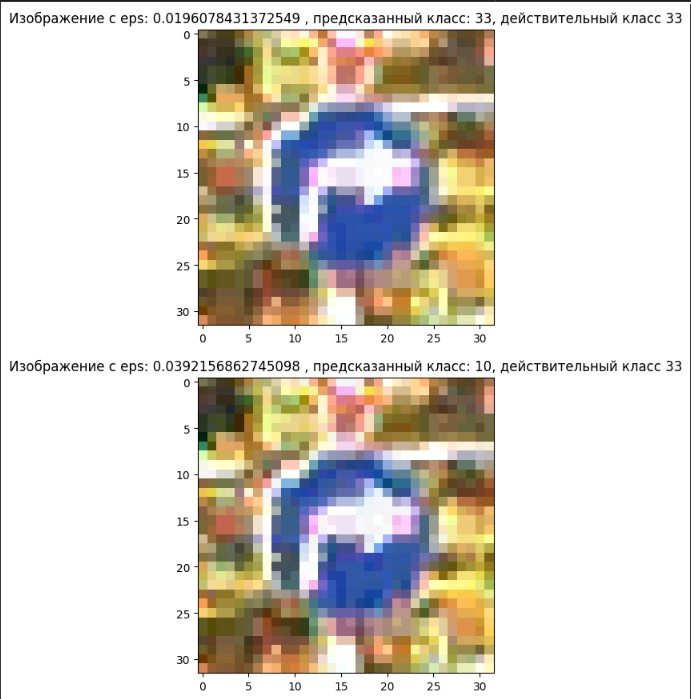
Далее нужно построить график зависимости точности предсказания модели на атакованных изображениях от параметра искажения.

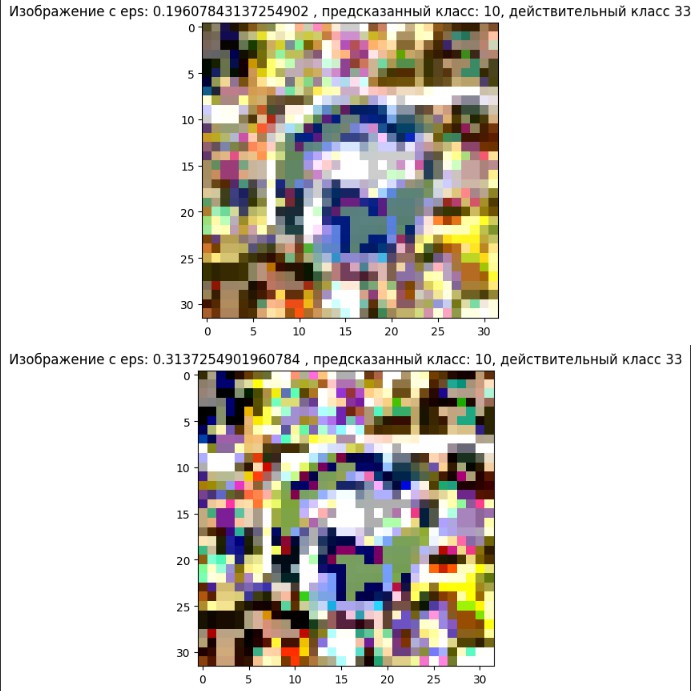
Исходя из графика можно сделать вывод, что данные методы имеют примерно одинаковую эффективность.



Теперь необходимо повторить эксперимент с атаками FGSM и PGD на базе модели VGG16. Ниже представлено отображение исходного и атакующих изображений для атаки FGSM







Далее нужно построить график зависимости точности предсказания модели на атакованных изображениях от параметра искажения.

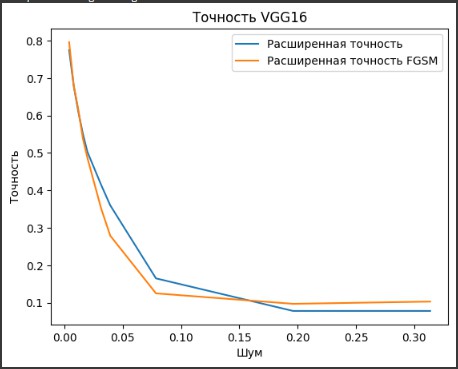


Таблица 2. Задание 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Модель** | **Исходные изображения, %** | **Adversarial images**  **𝜖=1/255, %** | **Adversarial images**  **𝜖=5/255, %** | **Adversarial images**  **𝜖=10/255, %** |
| ResNet50  FGSM | 91 | 72.9 | 32.4 | 15.8 |
| ResNet50  PGD | 91 | 70.3 | 31.4 | 20.7 |
| VGG16  FGSM | 93 | 79.6 | 48.7 | 27.9 |
| VGG16  PGD | 93 | 77.5 | 50.3 | 36 |

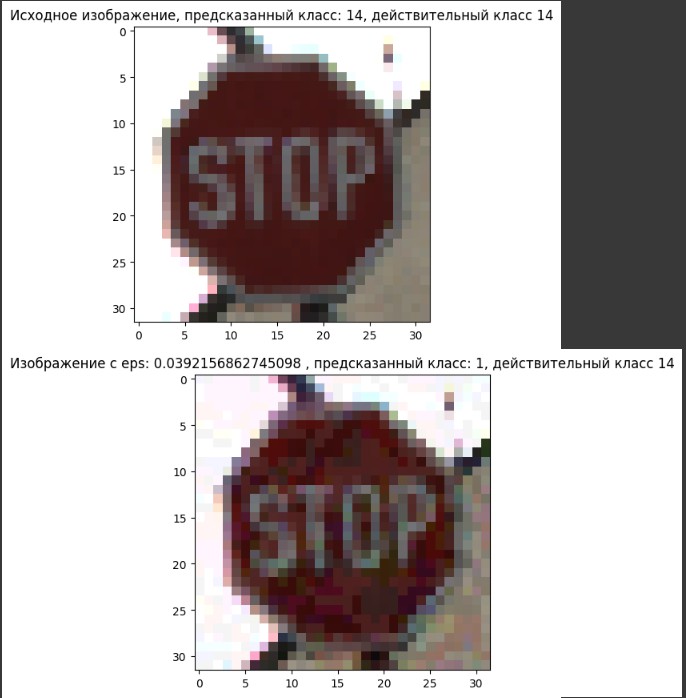
# Задание 3

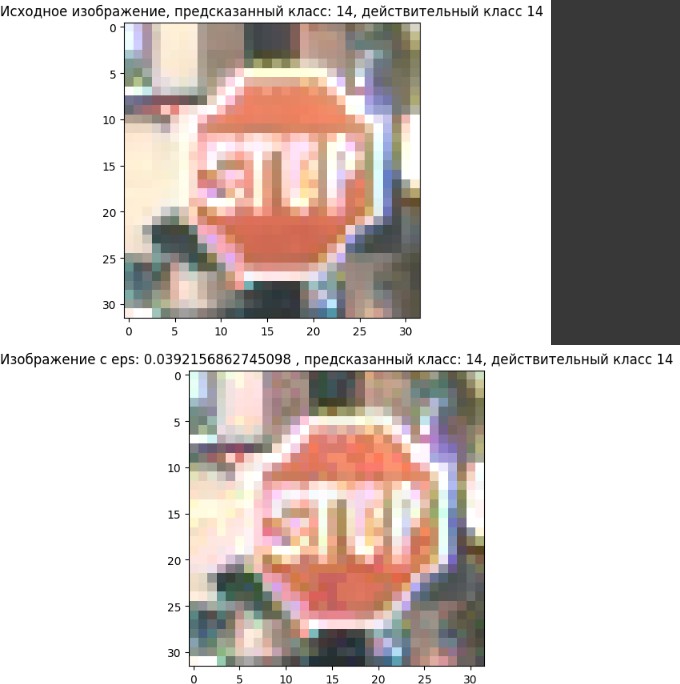
Применение целевой атаки уклонения методом белого против моделей глубокого обучения.

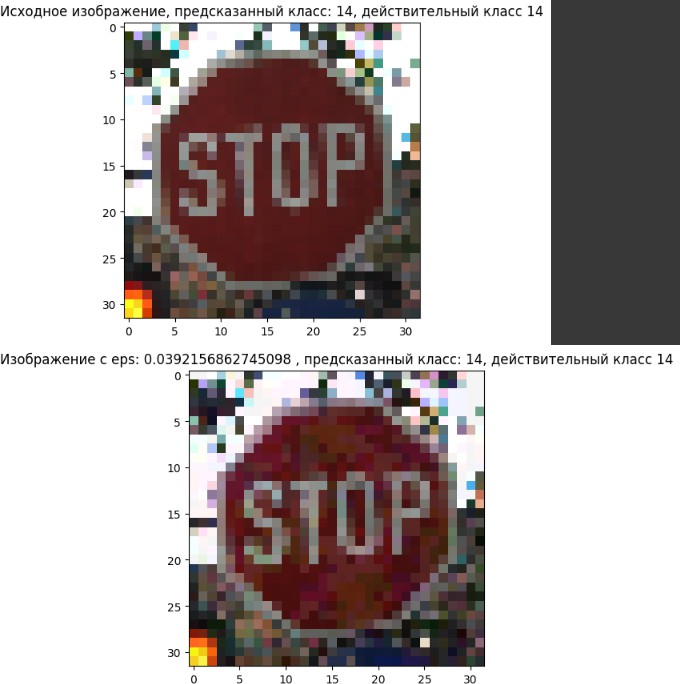
Шаг 1: Используйте изображения знака «Стоп» (label class 14) из тестового набора данных. Примените атаку PGD на знак «Стоп» с целью классификации его как знака «Ограничение скорости 30» (target label class = 1). Изменяйте значения искажений 𝜖𝜖= [1/255, 3/255, 5/255, 10/255, 20/255, 50/255, 80/255], и заполните отчёт значениями точности классификации изображений знаков "Стоп" и "Ограничение скорости 30".

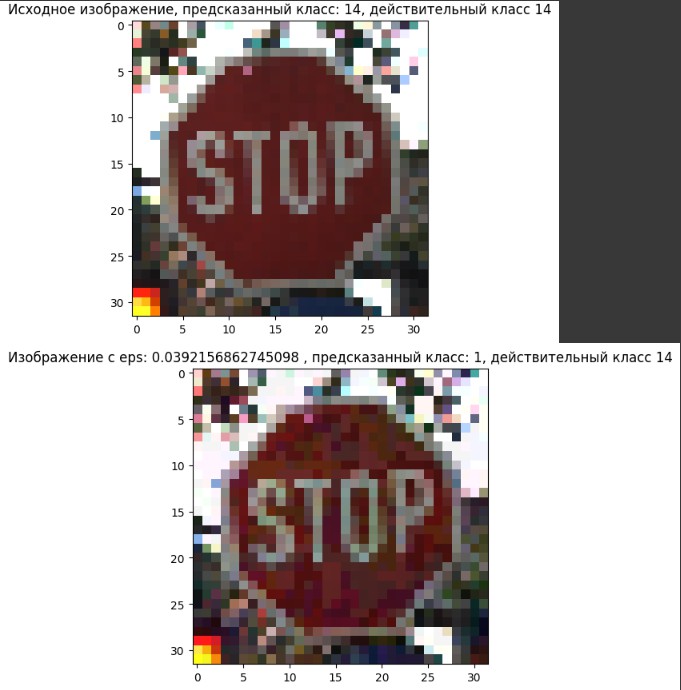
Шаг 2: Повторите атаку методом FGSM, и объясните производительность по сравнению с PGD. Сравните результаты атак PGD и FGSM между собой.

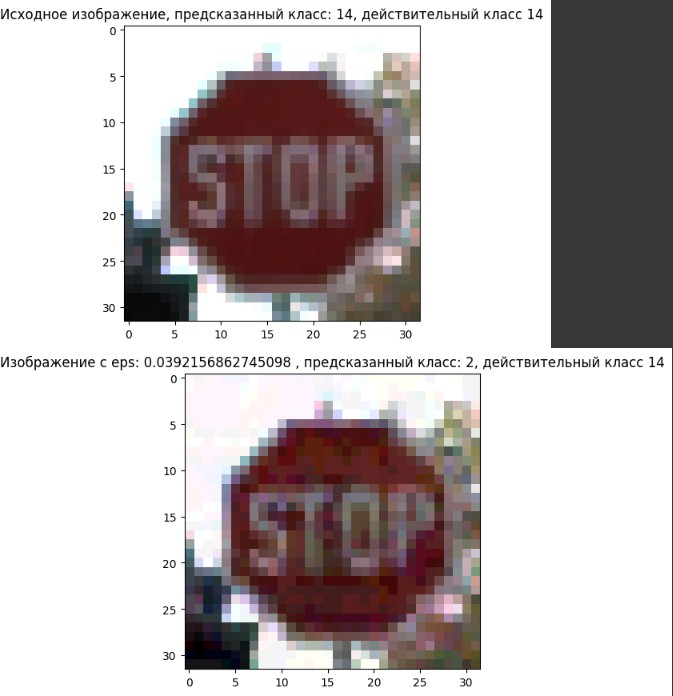
Шаг 1. Применение атаки PGD на знак «STOP» с целью классификации его как знака «Ограничение скорости 30».



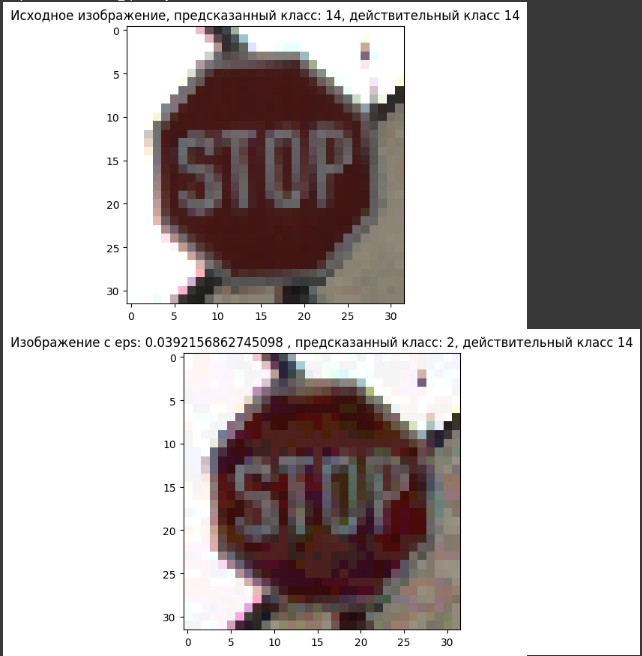


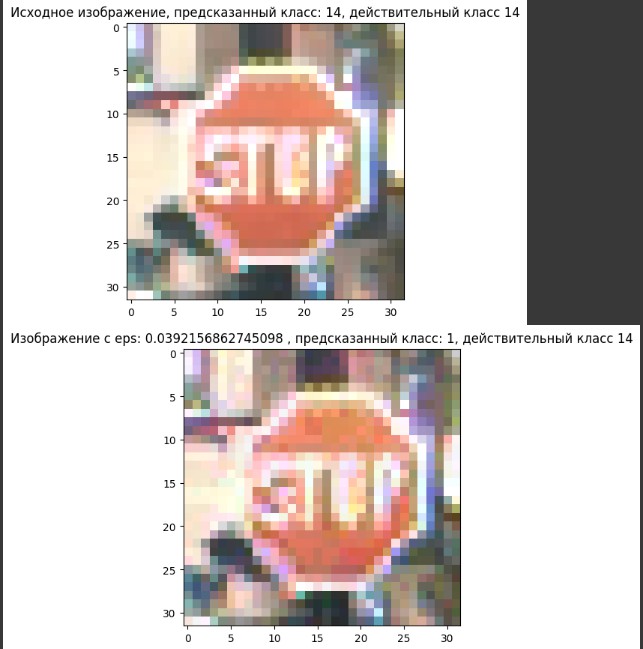


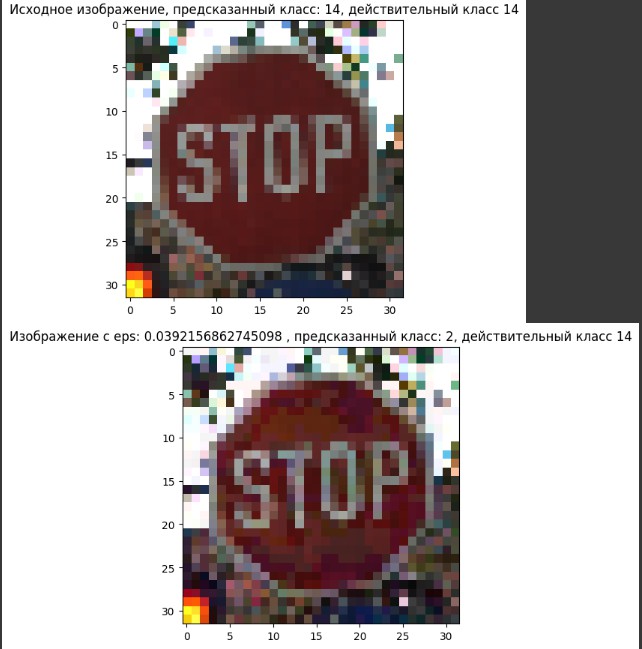


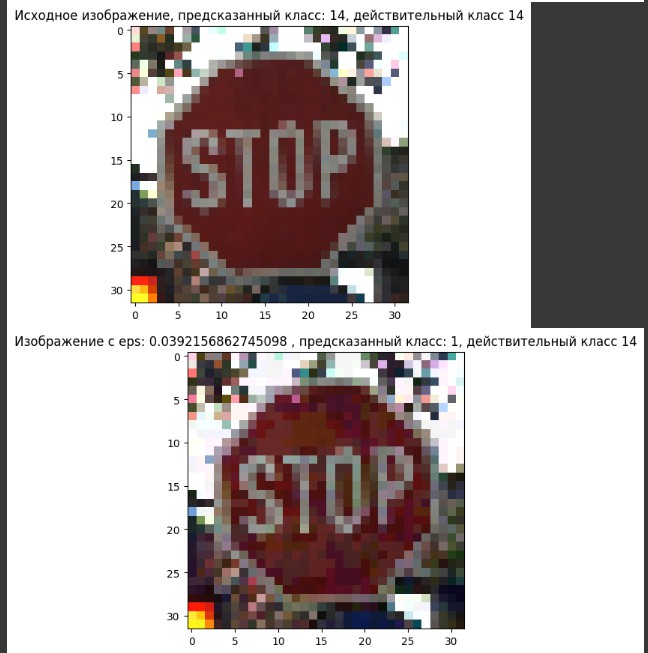


Повторение атаки методом FGSM









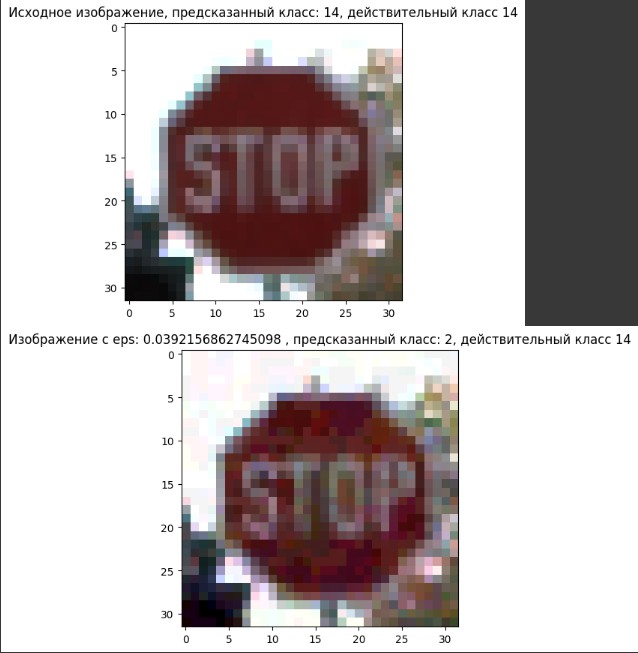


Таблица 3. Задание 3

|  |  |  |
| --- | --- | --- |
| **Искажение** | **PGD attack – Stop sign**  **images, %** | **FGSM attack – Stop**  **sign images, %** |
| 𝜖=1/255 | 97.4 | 88.1 |
| 𝜖=3/255 | 90 | 71.1 |
| 𝜖=5/255 | 79.6 | 47.8 |
| 𝜖=10/255 | 70.4 | 11.5 |
| 𝜖=20/255 | 37.8 | 0 |
| 𝜖=50/255 | 1.9 | 0 |
| 𝜖=80/255 | 0.3 | 0 |

# Вывод

Метод FGSM не подходит для целевых атак, так как с ростом eps и шума, классификация будет ошибочной. Наилучшим значением искажения будет являться 10/255, после этого модель будет ошибаться.

Метод PGD подходит для целевых атак, при больших значениях eps модель всегда будет определять заданный нами класс, однако минусом здесь является то, что изображение будет довольно сильно искажаться. Наилучшее значение искажения для данной атаки является 50/255