

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»

**Отчёт по лабораторной работе №2**

по дисциплине: «Анализ защищенности систем искусственного интеллекта»

Группа:

ББМО-02-22

Выполнила:

Бардасова И.А.

Проверил:

Спирин А.А.

Москва, 2023

Содержание

[Введение 3](#_Toc153941545)

[Ход выполнения работы 4](#_Toc153941546)

[Задание 1 5](#_Toc153941547)

[Задание 2 8](#_Toc153941548)

[Задание 3 17](#_Toc153941549)

[Заключение 25](#_Toc153941550)

# Введение

**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.

# Ход выполнения работы

**Шаг 1.** **Набор данных:** Для этой части используем набор данных GTSRB (German Traffic Sign Recognition Benchmark). Набор данных состоит примерно из 51 000 изображений дорожных знаков (рис. 2). Загрузим набор данных по ссылке:

<https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign> (рис. 1).

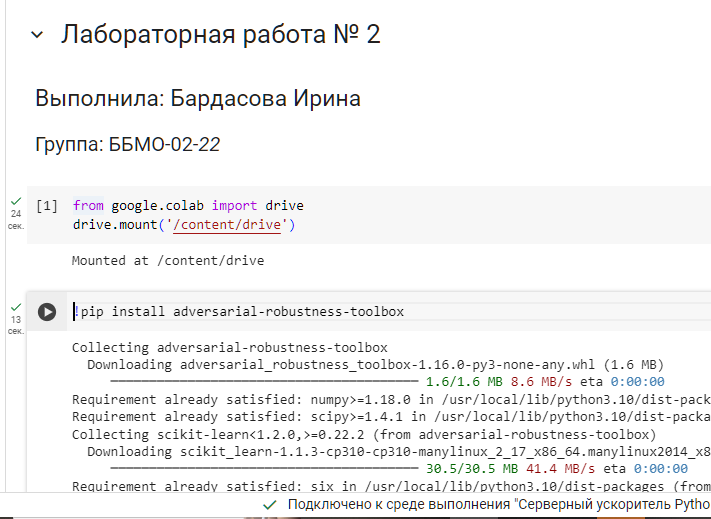


Рисунок 1 – Загрузка данных

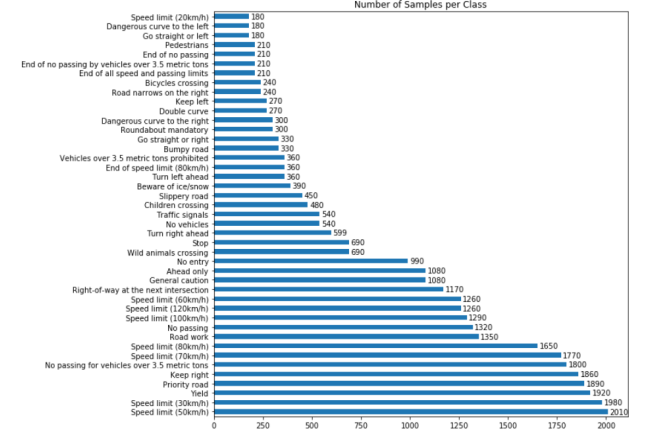


Рисунок 2 – Распределение изображений в GTRSB

# Задание 1

**Шаг 2.** **Обучить 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB.** Ресурсы колаба не безграничны, поэтому используем только часть набора данных. Использовали следующие модели нейронных сетей: ResNet50 и VGG16. Будем использовать необходимые фреймворки. Поделим набор данных на обучающую и тестовую в соотношении 70/30.

Сначала извлечем изображения для создания тренировочной выборки.

На выходе, мы получим матричное представление изображения (рис. 3).

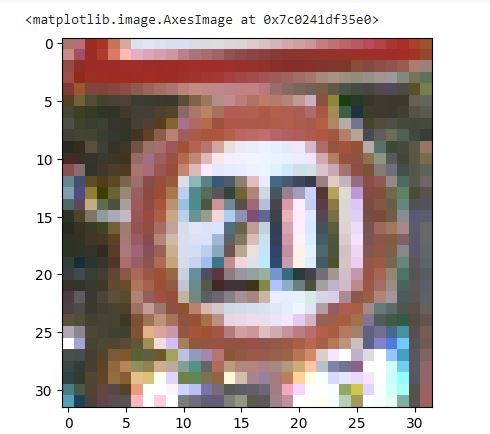


Рисунок 3 – Матричное представление изображения

**Шаг 3. Построение первой модели: ResNet50 (рис. 4).**

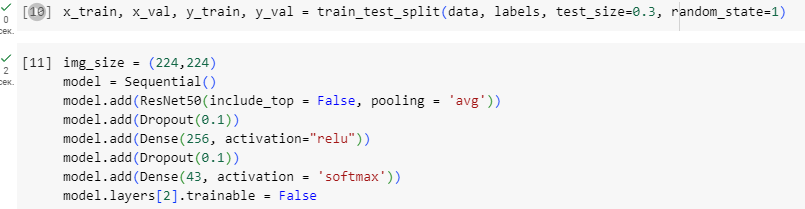


Рисунок 4 – Модель ResNet50

Определили оптимальное значение эпох обучения (5) и размера пакета (64). Для валидации будут выбраны 30 процентов тренировочного набора, сама валидация показана на рисунке 6. Графики процесса обучения представлены на рисунке 5.

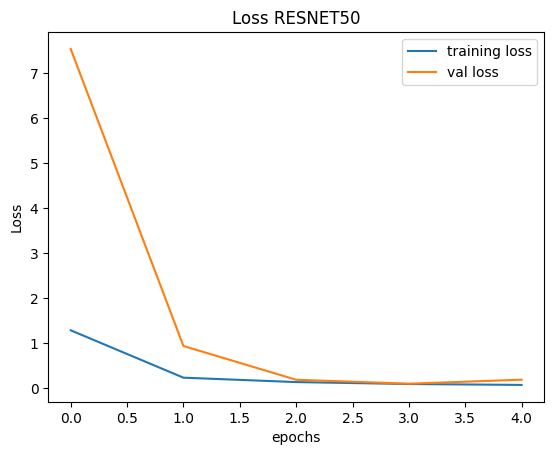
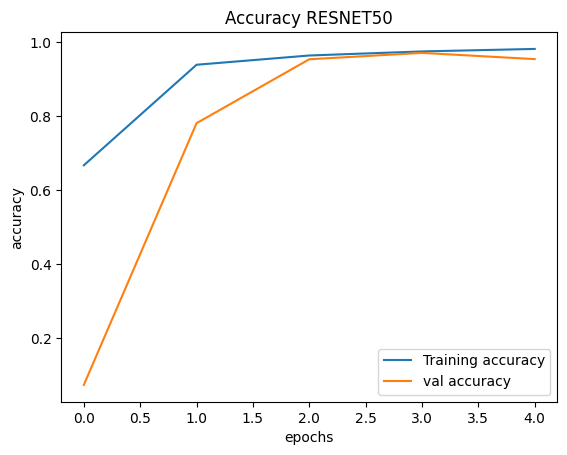


Рисунок 5 – Графики ResNet50



Рисунок 6 – Валидация ResNet50

Протестируем нашу модель на тестовом наборе. Результат валидации можно увидеть на рисунке 7. Итоговая точность составила – 90%.

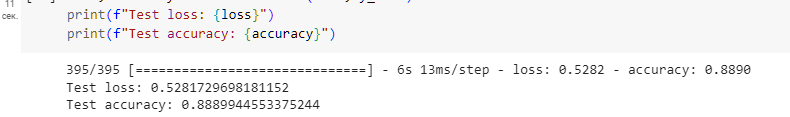


Рисунок 7 – Тестирование ResNet50

**Шаг 4. Построение второй модели: VGG16 (рис. 8).**



Рисунок 8 – Модель VGG16

Графики процесса обучения модели VGG16 показаны на рисунке 9. Валидационный результат представлен на рисунке 10.

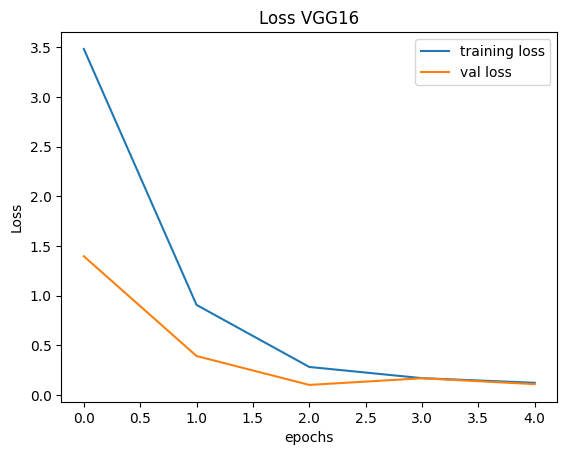
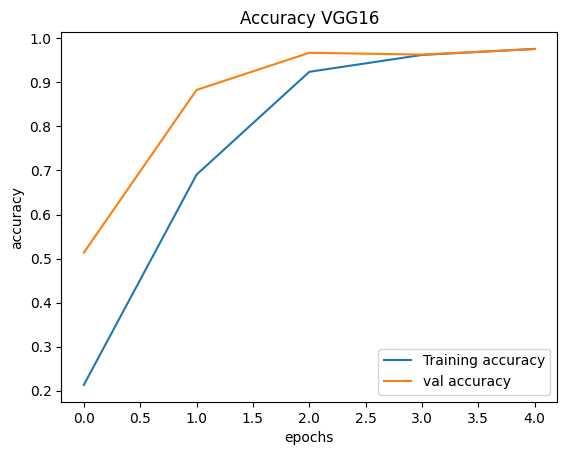


Рисунок 9 – Графики VGG16



Рисунок 10 – Валидация VGG16

Тестирование обученной модели на валидационном наборе (рис. 11):

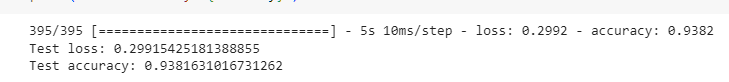


Рисунок 11 – Тестирование VGG16

**Шаг 5. Результаты:** Подведём результаты по первому заданию в таблице 1.1.

Таблица 1.1 – Результаты

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Модель** | **Обучение** | | **Валидация** | | **Тест** | |
| **loss** | **accuracy** | **loss** | **accuracy** | **loss** | **accuracy** |
| **ResNet50** | 0,0758 | 0,9815 | 0,1925 | 0,9539 | 0,5282 | 0,8890 |
| **VGG16** | 0,1215 | 0,9760 | 0,108 | 0,9758 | 0,2992 | 0,9382 |

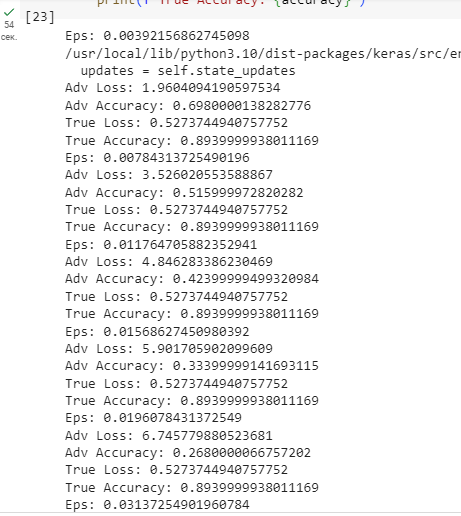
# Задание 2

**Шаг 6. Применим нецелевую атаку уклонения на основе белого ящика против моделей глубокого обучения.** Реализуем атаку Fast Gradient Sign Method (FGSM) и Projected Gradient Descent (PGD).

Для создания нецелевых атакующих примеров используем первые 1,000 изображений из тестового множества. Также используем следующие значения параметра искажения для атак на изображения: 𝜖 = [1/255, 2/255, 3/255, 4/255, 5/255, 8/255, 10/255, 20/255, 50/255, 80/255].

**Шаг 7.** **Атака FGSM на ResNet50.** Создаем модель атаки, которая будет основываться на обученном классификаторе для внесения шума в изображение.

Результаты каждого параметра на рисунке 12.



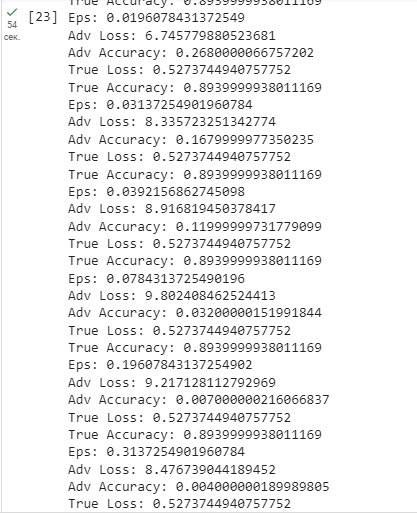


Рисунок 12 – Параметры искажения для атак на изображения

График зависимости точности предсказания модели на атакованных изображениях от параметра искажения представлена на рисунке 13.

Отобразим исходное изображение из датасета и атакующие изображения с указанием величины параметра 𝜖 = [1/255, 5/255, 10/255, 50/255, 80/255] (рис. 14).

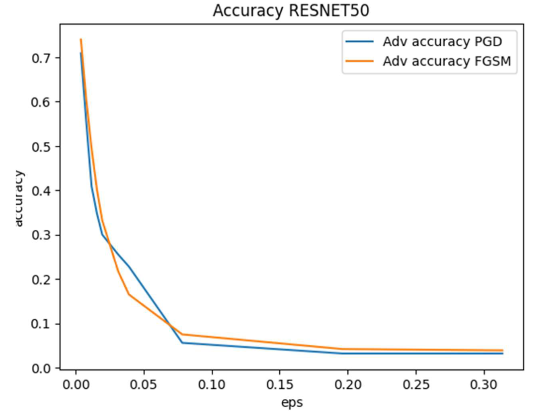
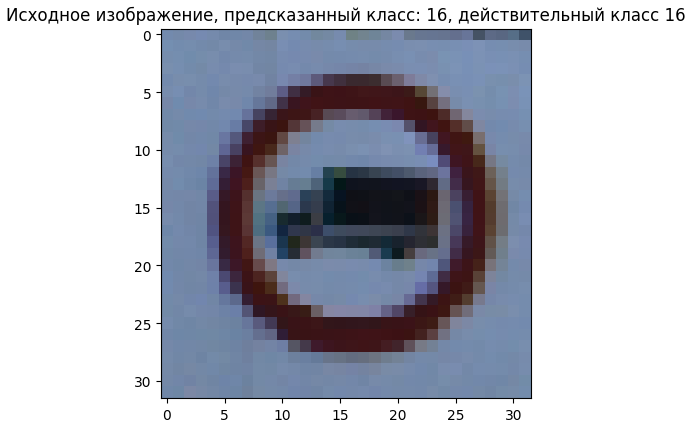
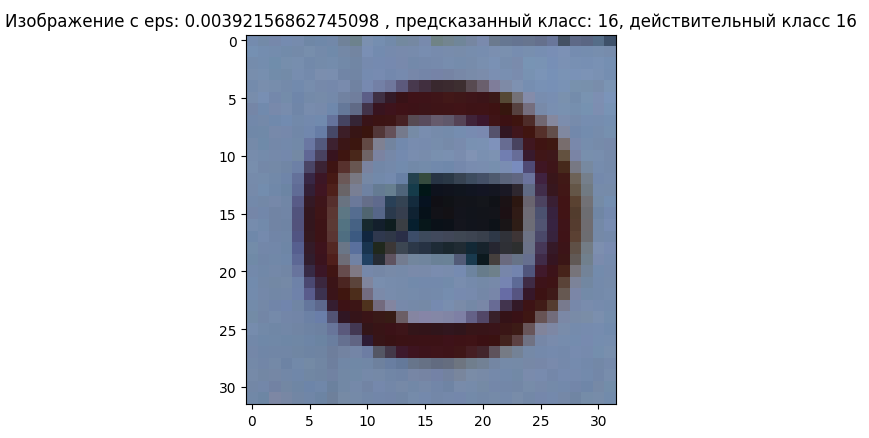
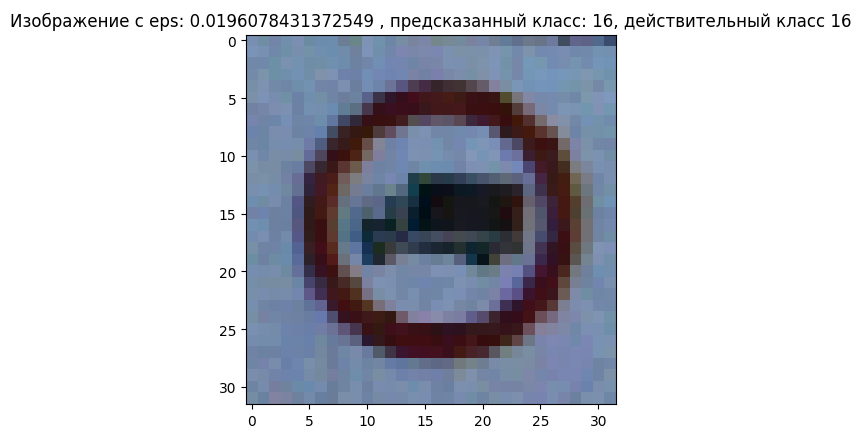
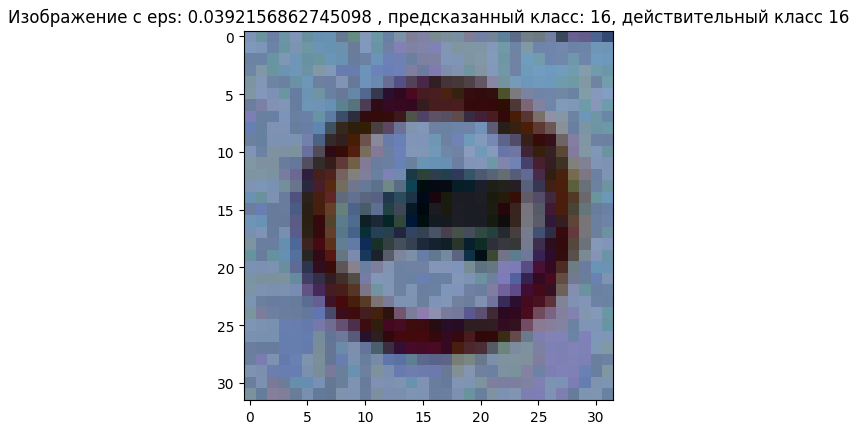


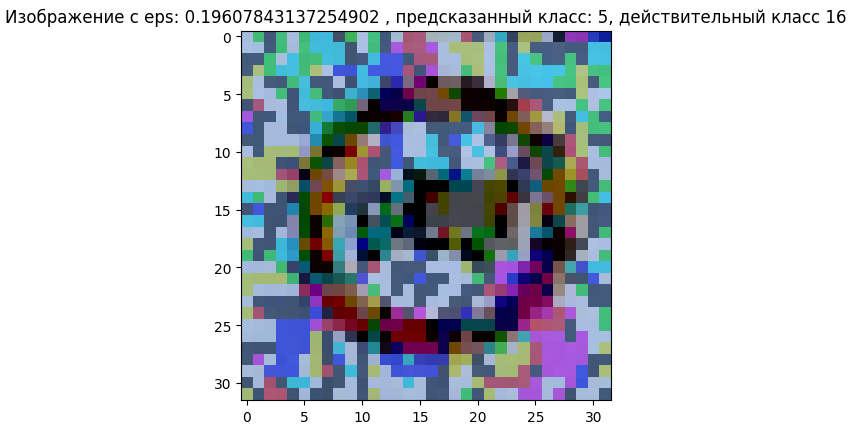
Рисунок 13 – График зависимости











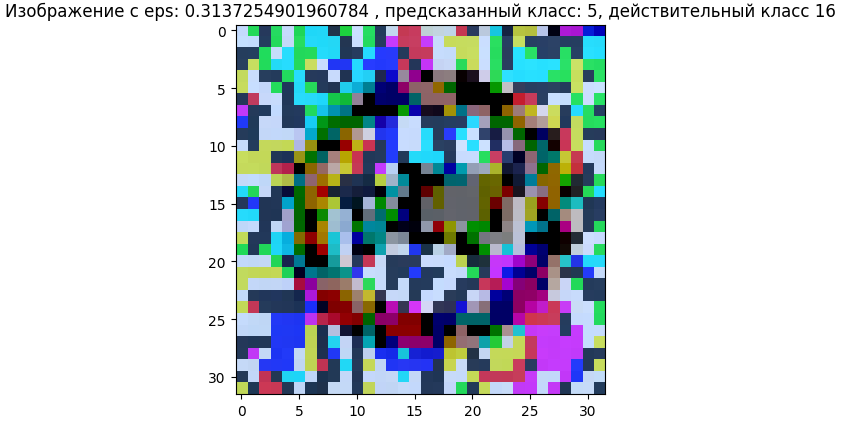
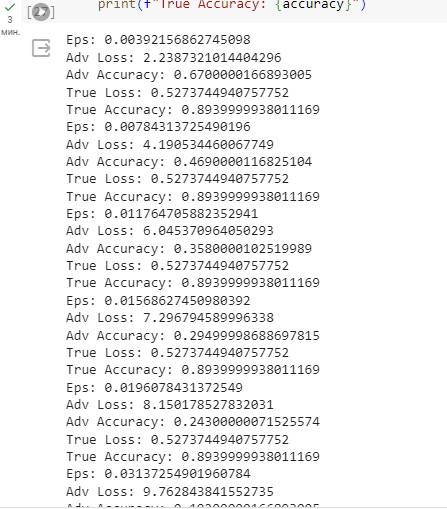


Рисунок 14 – Исходные изображения и искажённые изображения FGSM

**Шаг 8.** Теперь сделаем то же самое с ResNet50 через PGD. Подобно FGSM реализуем атаку PGD для различных значений eps (рис. 16). Результаты каждого параметра на рисунке 15.



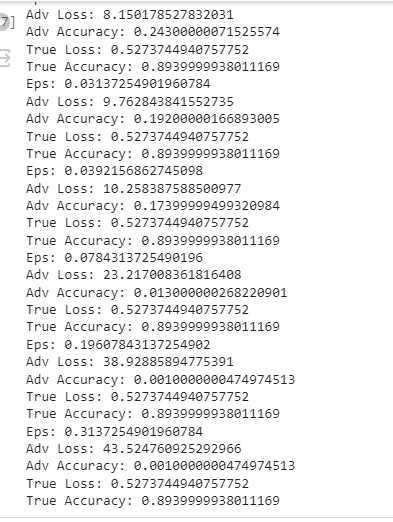


Рисунок 15 – Параметры искажения для атак на изображения

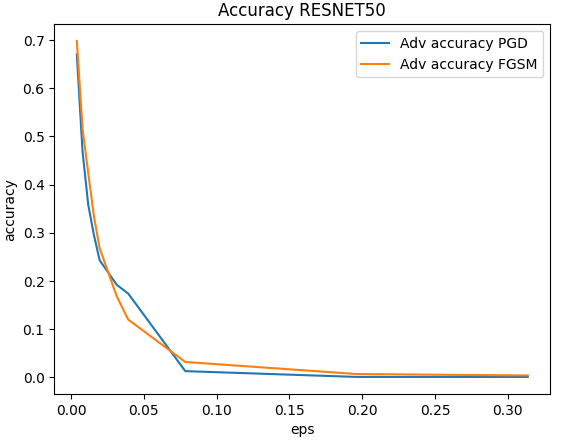
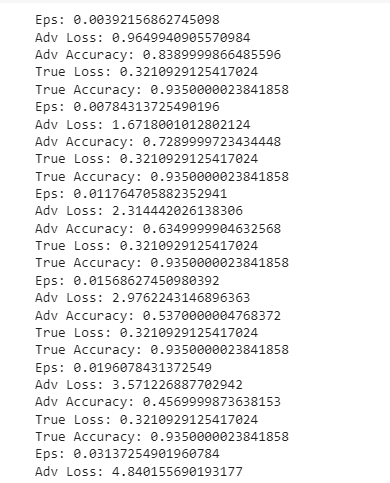


Рисунок 16 – График зависимости PGD

**Шаг 9. Сделаем то же самое для VGG16 – начнём с атаки FGSM (рис. 18).** Результаты каждого параметра на рисунке 17.



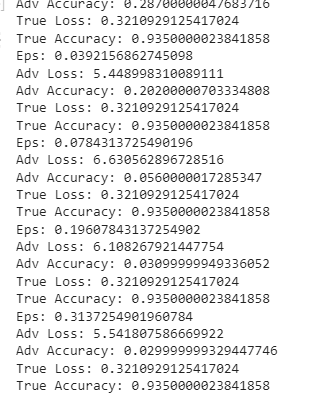
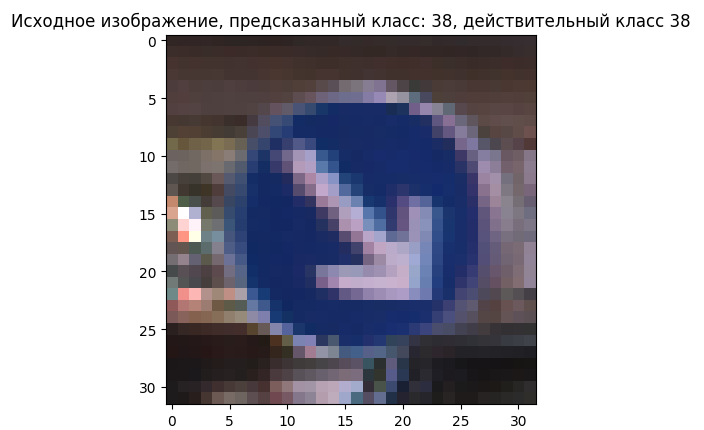
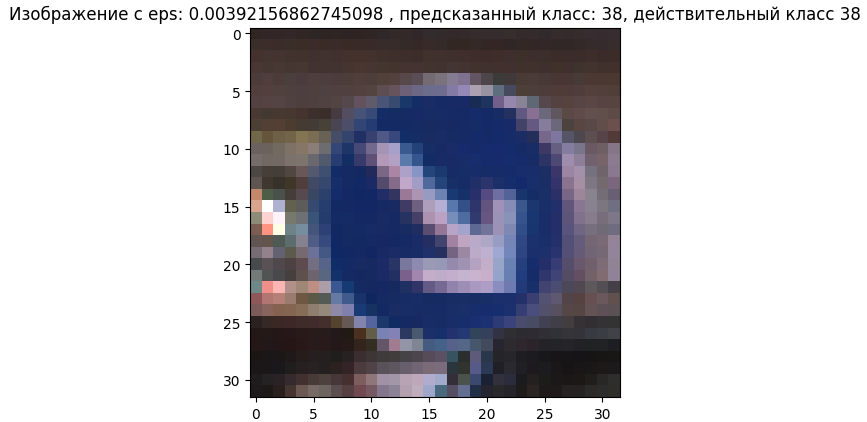
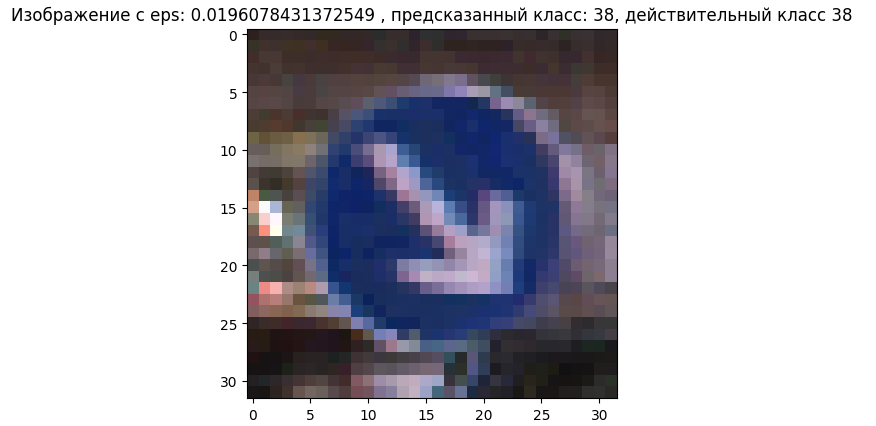
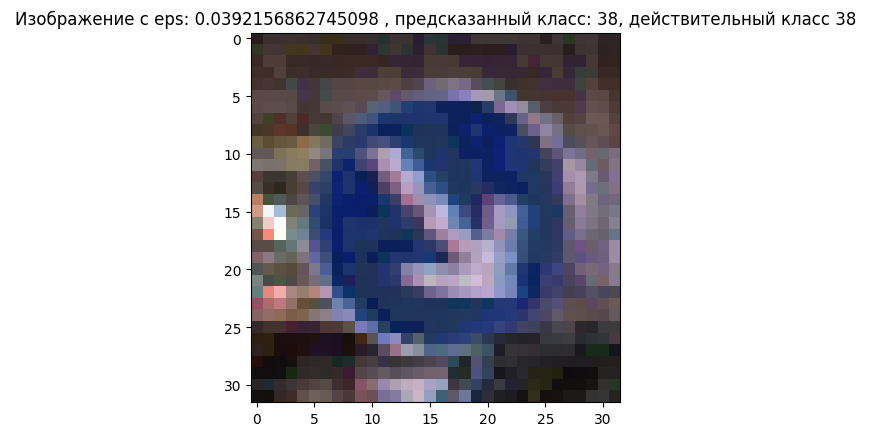


Рисунок 17 – Параметры искажения для атак на изображения











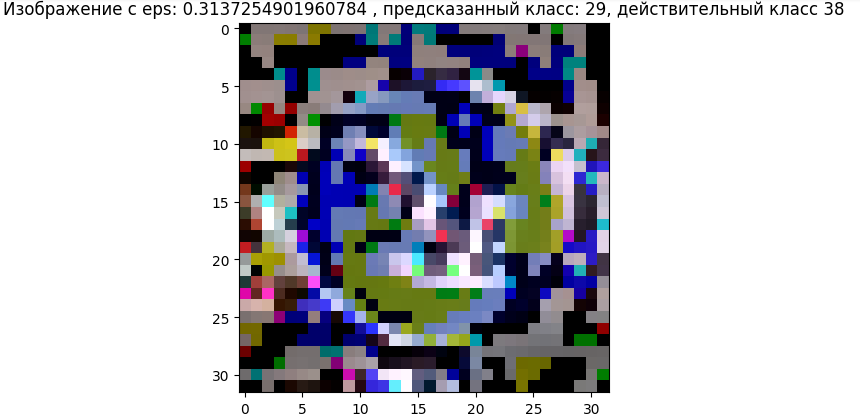
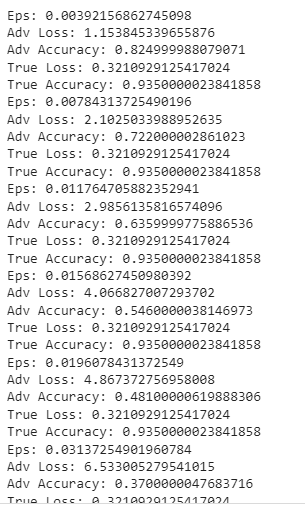


Рисунок 18 – Исходные изображения и искажённые изображения FGSM

**Шаг 10. VGG16 PGD (рис. 20).** Подобно FGSM реализуем атаку PGD для различных значений eps. Результаты каждого параметра на рисунке 19.



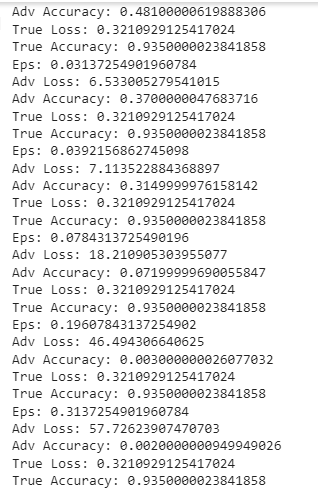


Рисунок 19 – Параметры искажения для атак на изображения

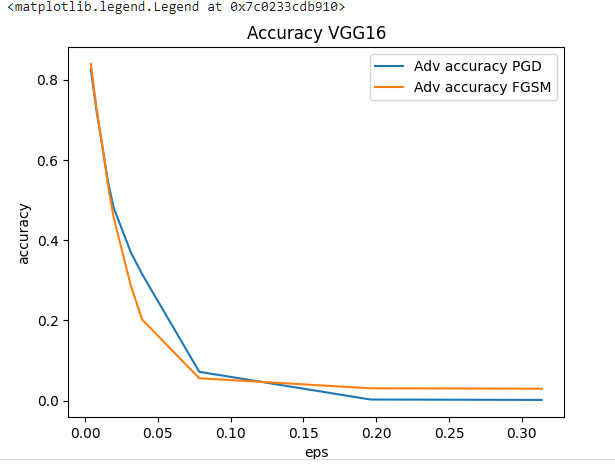


Рисунок 20 – График зависимостей PGD

**Шаг 11. Результаты (таблица 2.1).**

Таблица 2.1 - Результаты

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Исходные изображения | Adversarial images 𝜖𝜖=1/255 | Adversarial images 𝜖𝜖=5/255 | Adversarial images 𝜖𝜖=10/255 |
| ResNet50 - FGSM | loss: 0.5275  accuracy: 0.894 | loss: 1.9604  accuracy: 0.698 | loss: 6.7457  accuracy: 0.268 | loss: 8.4767  accuracy: 0.004 |
| ResNet50 - PGD | loss: 0.5275  accuracy: 0.894 | loss: 2.2387  accuracy: 0.67 | loss: 8.1501  accuracy: 0.243 | loss: 43.5247  accuracy: 0.001 |
| VGG16 - FGSM | loss: 0.3210  accuracy: 0.935 | loss: 0.9649  accuracy: 0.839 | loss: 3.5712  accuracy: 0.457 | loss: 5.5418  accuracy: 0.0299 |
| VGG16 - PGD | loss: 0.3210  accuracy: 0.935 | loss: 1.1538  accuracy: 0.825 | loss: 4.8673  accuracy: 0.4810 | loss: 57.7262  accuracy: 0.3210 |

# Задание 3

Применение целевой атаки уклонения методом белого против моделей глубокого обучения.

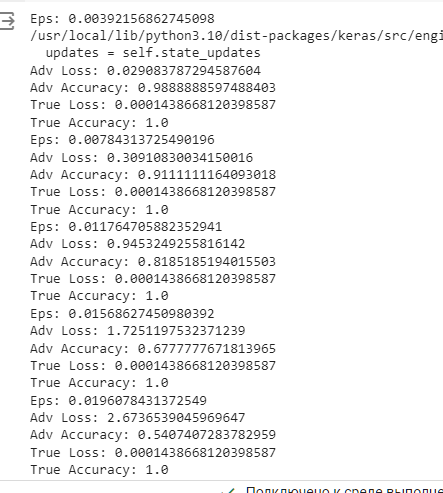
**Шаг 12.** Используем изображения знака «Стоп» (label class 14) из тестового набора данных. Применим атаку Projected Gradient Descent (PGD) на знак «Стоп» с целью классификации его как знака «Ограничение скорости 30» (target label class = 1). Будем изменять значения искажений 𝜖𝜖= [1/255, 3/255, 5/255, 10/255, 20/255, 50/255, 80/255].

Повторим атаку методом FGSM и заполним таблицу 3.1.

Таблица 3.1 – Результаты

|  |  |  |
| --- | --- | --- |
| **𝜖** | **FGSM - Stop** | **PGD – Stop** |
| 𝜖 = 1/255 | loss: 0.0290  accuracy: 0.9888 | loss: 0.0308  accuracy: 0.9888 |
| 𝜖 = 3/255 | loss: 0.9453  accuracy: 0.8185 | loss: 0.5882  accuracy: 0.8555 |
| 𝜖 = 5/255 | loss: 2.6736  accuracy: 0.5407 | loss: 1.399  accuracy: 0.6296 |
| 𝜖 = 10/255 | loss: 7.3347  accuracy: 0.088 | loss: 2.5267  accuracy: 0.47037 |
| 𝜖 = 20/255 | loss: 10.321  accuracy: 0.0 | loss: 8.7211  accuracy: 0.1296 |
| 𝜖 = 50/255 | loss: 10.1705  accuracy: 0.0 | loss: 21.8093  accuracy: 0.0 |
| 𝜖 = 80/255 | loss: 8.8265  accuracy: 0.0 | loss: 23.3313  accuracy: 0.0 |

Подробнее ниже. Результаты каждого параметра **FGSM** представленына рисунке 21.



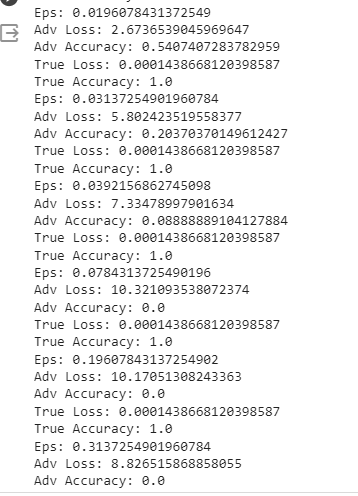


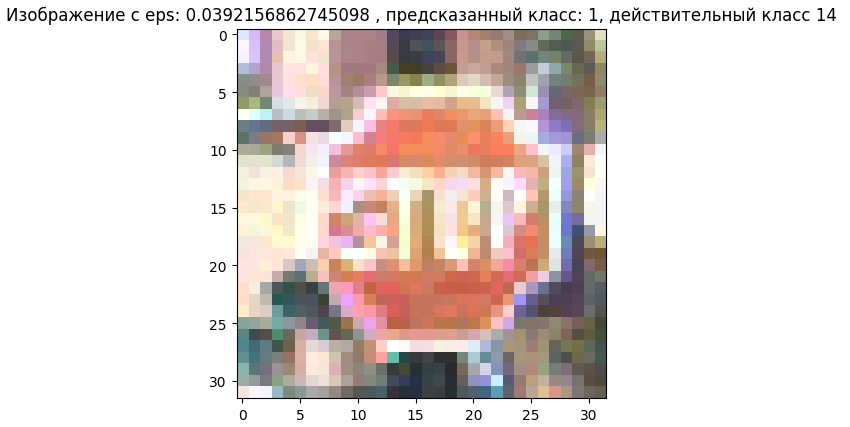
Рисунок 21 – Параметры FGSM

Выведем 5 примеров классификации класса 14 как класс 1 при помощи целевой FGSM атаки (рис. 22):



















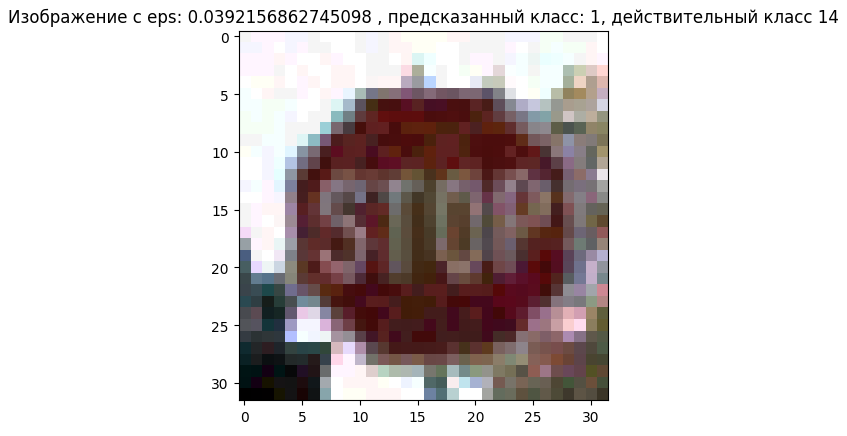
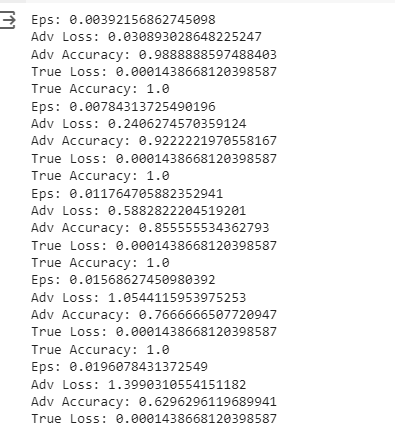


Рисунок 22 – FGSM искажение

Результаты каждого параметра **PGD** представленына рисунке 23.



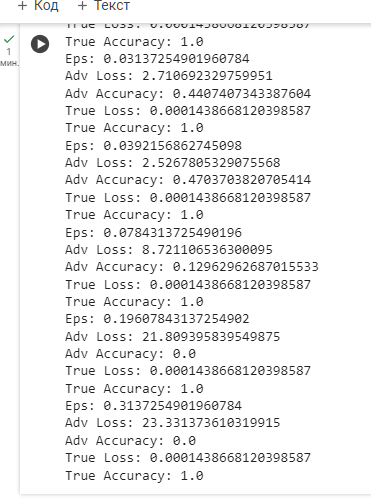
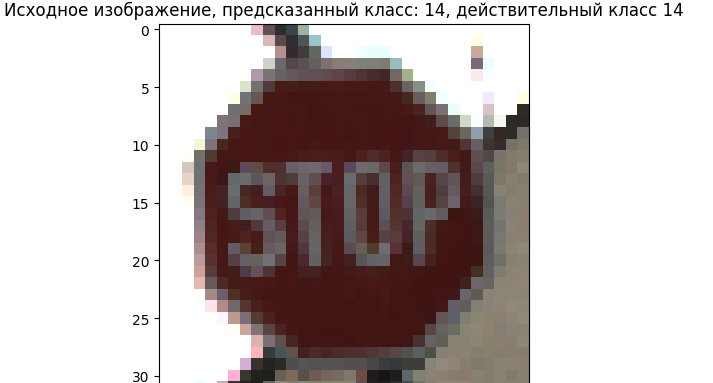
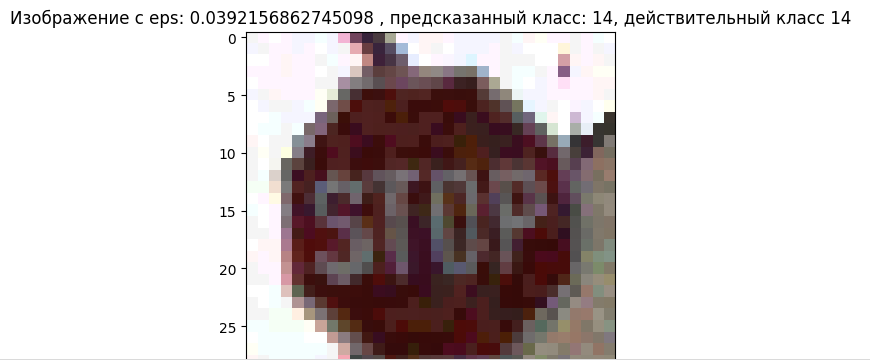


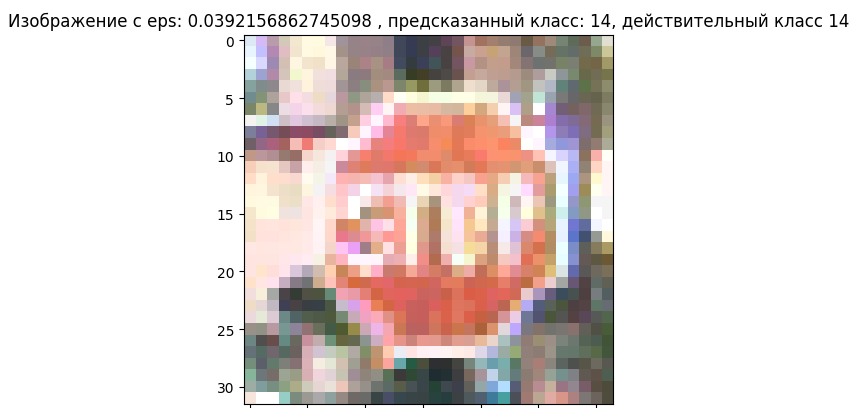
Рисунок 23 – Результаты для PGD

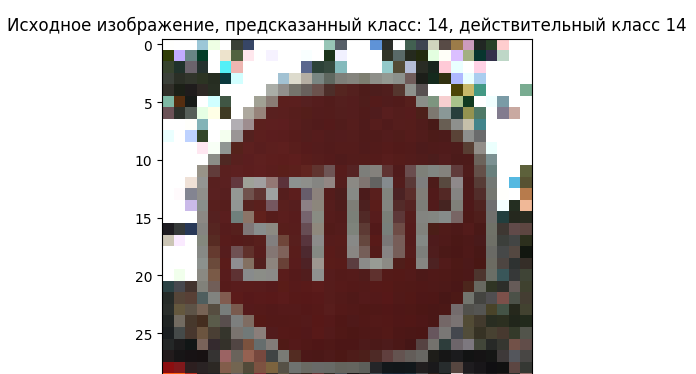
Выведем 5 примеров классификации класса 14 как класс 1 при помощи целевой PGD атаки (рис. 24):













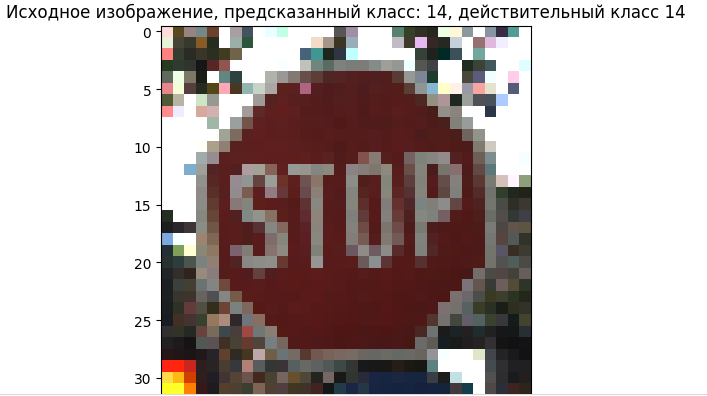








Рисунок 24 – PGD искажение

# Заключение

В результате выполнения работы были также проведены эксперименты по атаке на модели машинного обучения методом черного и белого ящика, а также целевые и нецелевые.

Также были рассмотрены модели VGG16 и ResNet50, VGG16 показала несколько большую устойчивость к атакам, хоть и не значительную, но уже не в рамках погрешности. Было отмечено сильно ухудшение качества и точности моделей по достижении отметки искажения в 20/255. Метод FGSM плохо подходит для целевых атак. С ростом искажения классификация начинает давать сбои. PGD больше подходит для целевых атак. При больших искажениях, модель будет определять нужный класс, но картинка сильно испортится шумом.