|  |
| --- |
|  |
| 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 №2

по дисциплине: «Анализ защищенности систем искусственного интеллекта»

**Выполнил**:

Студент группы ББМО-01-22

Загороднов Егор Алексеевич

**Проверил**:

К.т.н. Спирин Андрей Андреевич

Москва 2023

Содержание

[Задание 1 3](#_Toc151024349)

[Задание 2 7](#_Toc151024350)

[Задание 3 14](#_Toc151024351)

[Вывод 22](#_Toc151024352)

[Заключение 23](#_Toc151024353)

# Задание 1

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

В данном задании нам требуется обучить 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB. В качестве исходных данных была взята лишь часть набора примеров.

Набор данных был поделен на обучающую и тестовую выборки в соотношении 70 к 30.

Первая модель построена на базе ResNet50 и состоит из слоев, представленных на рисунке 1.

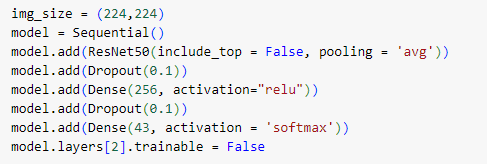


Рисунок 1 – Модель ResNet50

В результате множества перезапусков обучения, было определено оптимальное значение epochs и batch\_size. Для валидации будут выбраны 30 процентов тренировочного набора. При этом, история обучения будет сохранена, для последующего создания графиков. Валидация обучения показаны на рисунке 2.

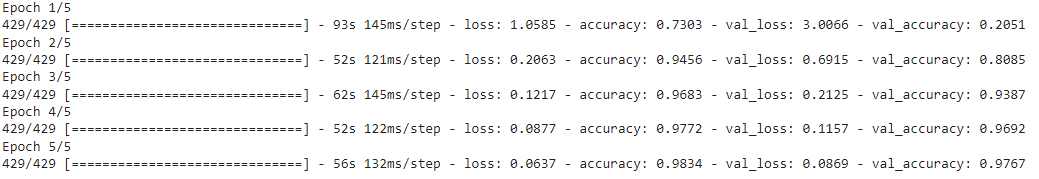


Рисунок 2 – Валидация ResNet50

Графики процесса обучения представлены на рисунках 3 и 4.

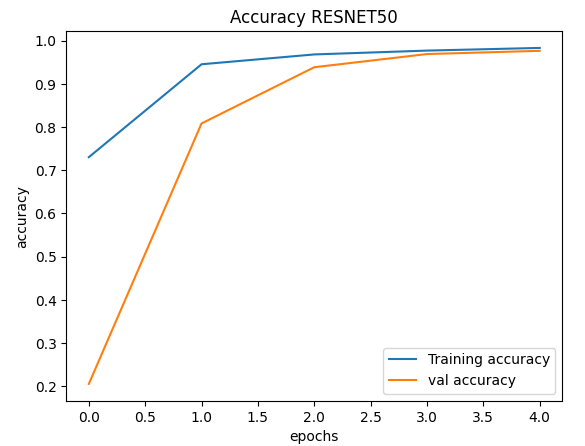


Рисунок 3 – Графики ResNet50

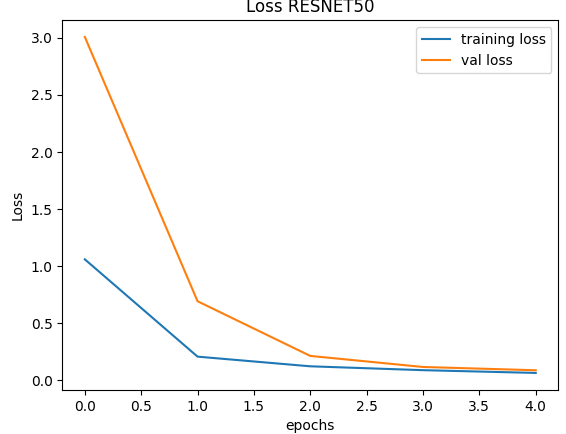


Рисунок 4 – Графики ResNet50

После обучения модель была протестирована на тестовом наборе. Видим, что итоговая точность составила 90 процентов. Показатели валидации приведены на рисунке 5.

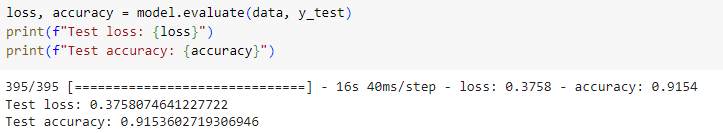


Рисунок 5 – Тестирование ResNet50

Вторая модель построена на базе VGG16. Загрузим уже готовый набор данных для тренировки. Модель VGG16 показана на рисунке 6.

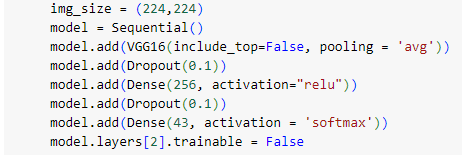


Рисунок 6 – Модель VGG16

Графики процесса обучения представлены на рисунках 7 и 8.

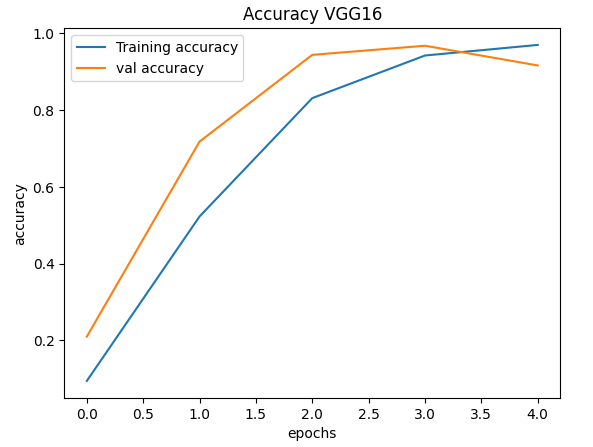


Рисунок 7 – Графики VGG16

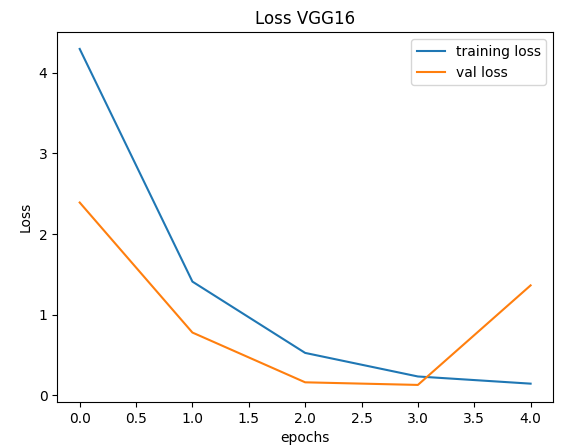


Рисунок 8 – Графики VGG16

Валидация показана на рисунке 9.

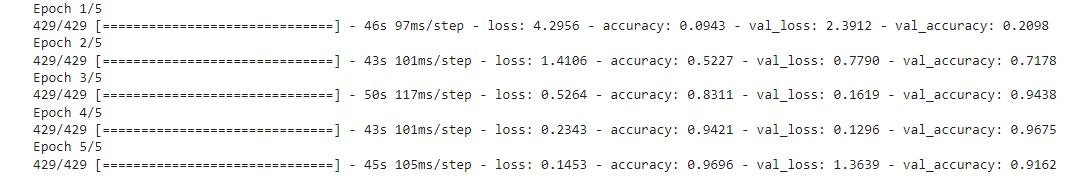


Рисунок 9 – Валидация VGG16

После обучения модель была протестирована на тестовом наборе Показатели проведенного тестирования приведены на рисунке 10.

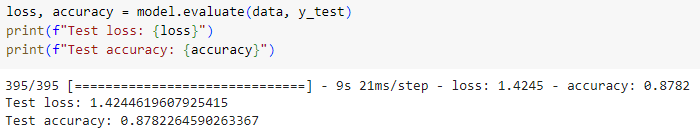


Рисунок 10 – Тестировние VGG16

Таблица с результатами по заданию представлена ниже.

Таблица 1 – Таблица по результатам

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Обучение | Валидация | Тест |
| ResNet50 | loss: 0.0637 accuracy: 0.9834 | loss: 0.0869 accuracy: 0.9767 | loss: 0.3758 accuracy: 0.9154 |
| VGG16 | loss: 0.1453 accuracy: 0.9696 | loss: 1.3639 accuracy: 0.9162 | loss: 1.4244 accuracy: 0.8782 |

# Задание 2

**Задание 2. Применить нецелевую атаку уклонения на основе белого ящика против моделей глубокого обучения.**

В данном задании требуется применить нецелевую атаку уклонения на основе белого ящика против моделей глубокого обучения.

Используем атаку FSGM и PGD на базе ResNet50 для создания нецелевых атакующих примеров используя первые 1,000 изображений из тестового множества.

Атаки на изображения проводятся со следующими параметрами искажения [1/255, 2/255, 3/255, 4/255, 5/255, 8/255, 10/255, 20/255, 50/255, 80/255].

График зависимости точности предсказания модели на атакованных изображениях от параметра искажения приведен на рисунке 11.

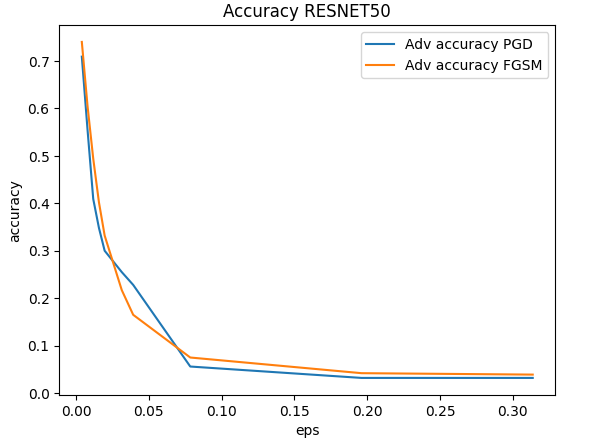


Рисунок 11 - График ResNet50

Для атаки FGSM, отобразим исходное изображение из датасета и атакующие изображения с указанием величины параметра 𝜖 = [1/255, 5/255, 10/255, 50/255, 80/255] . Изображение из датасета и атакующие изображения с указанием величины параметра 𝜖показан на рисунках 12-17.

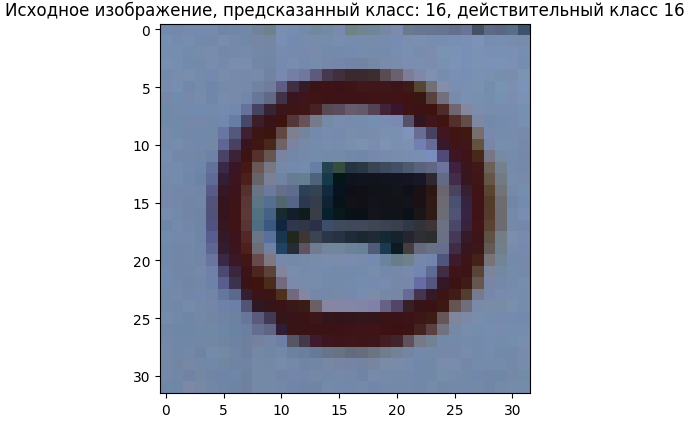


Рисунок 12 - Исходное изображение ResNet50

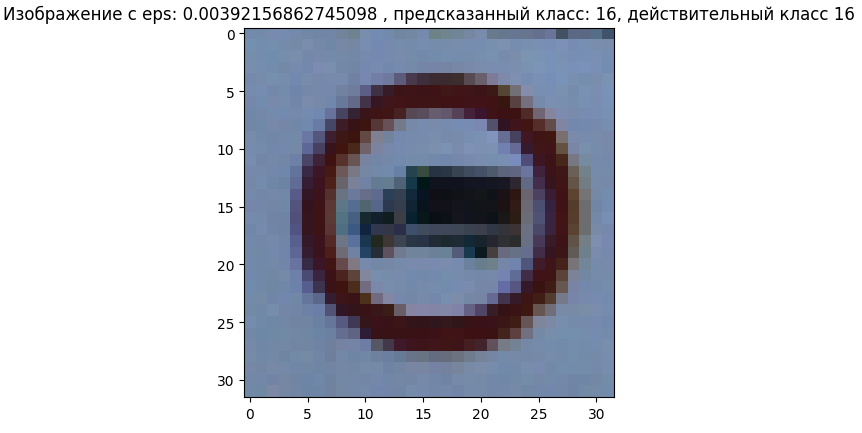


Рисунок 13 - Искаженное изображение ResNet50

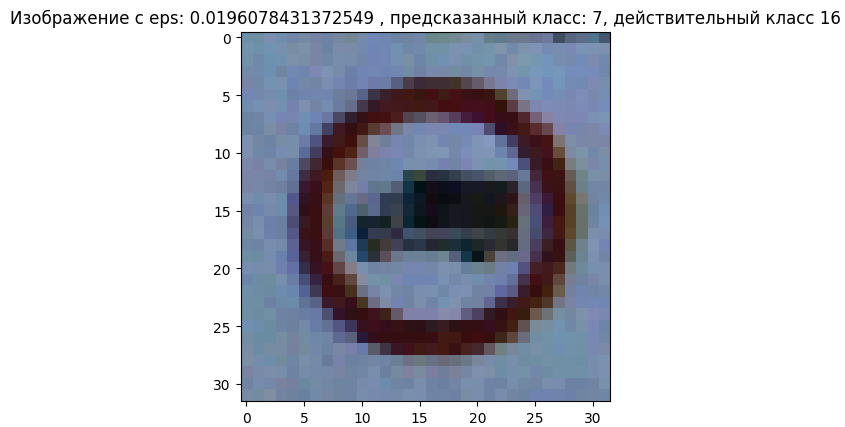


Рисунок 14 - Искаженное изображение ResNet50

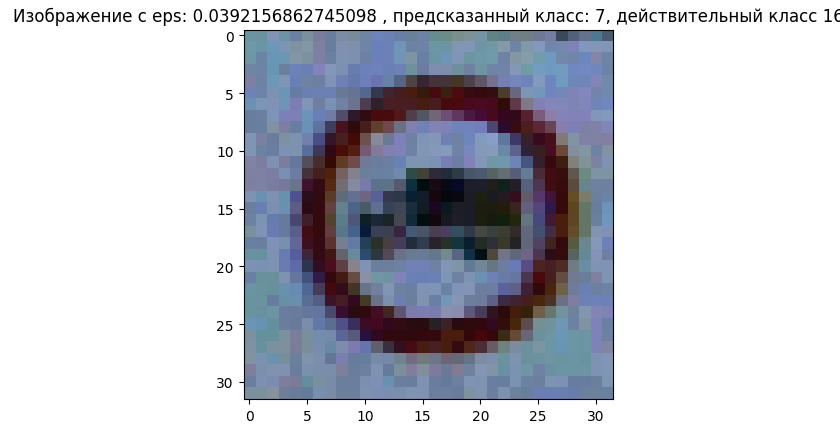


Рисунок 15 - Искаженное изображение ResNet50

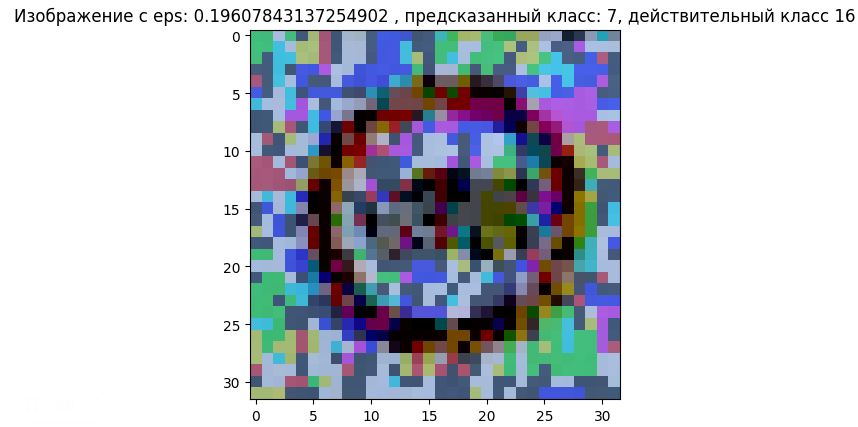


Рисунок 16 - Искаженное изображение ResNet50

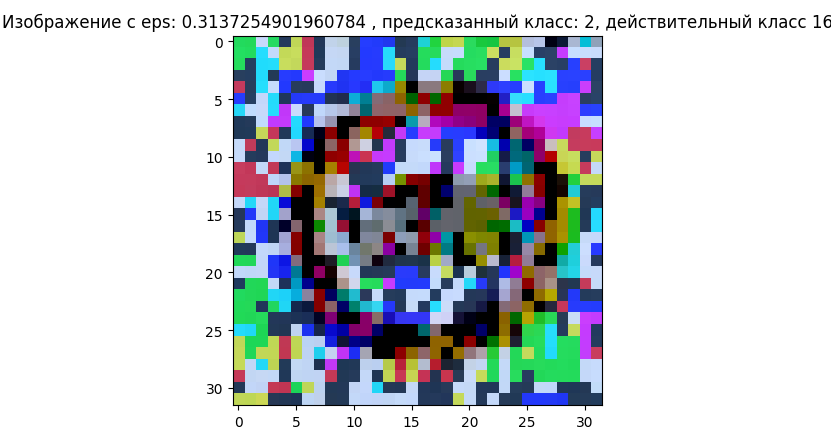


Рисунок 17 - Искаженное изображение ResNet50

Из этих всех изображений можем сделать вывод, что при росте eps, шум на картинке сильно увеличивается, и с 5/255 уже становится более заметен.

Повторим эксперимент с атаками FSGM и PGD на модель на базе VGG16.

График зависимости точности предсказания модели на атакованных изображениях от параметра искажения приведен на рисунке 18.

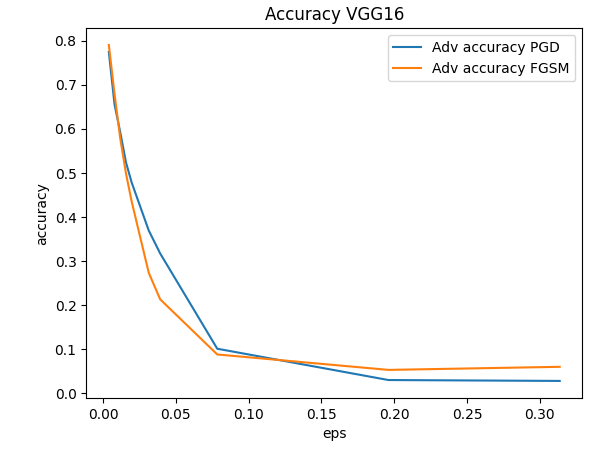


Рисунок 18 – График VGG16

Для атаки FGSM, отобразим исходное изображение из датасета и атакующие изображения с указанием величины параметра 𝜖 = [1/255, 5/255, 10/255, 50/255, 80/255]. Исходное изображение из датасета и атакующие изображения с указанием величины параметра 𝜖 показаны на рисунках 19-24

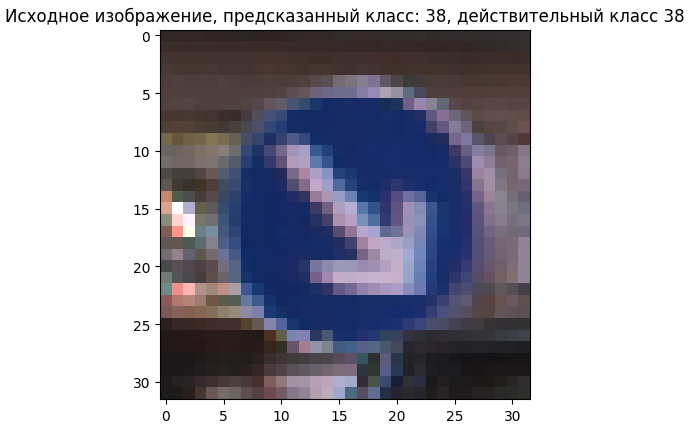


Рисунок 19 – Исходное изображение VGG16

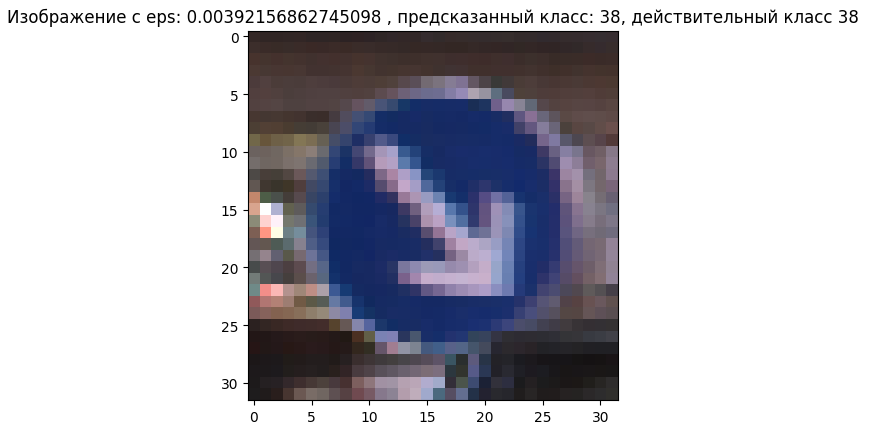


Рисунок 20 – Искаженное изображение VGG16

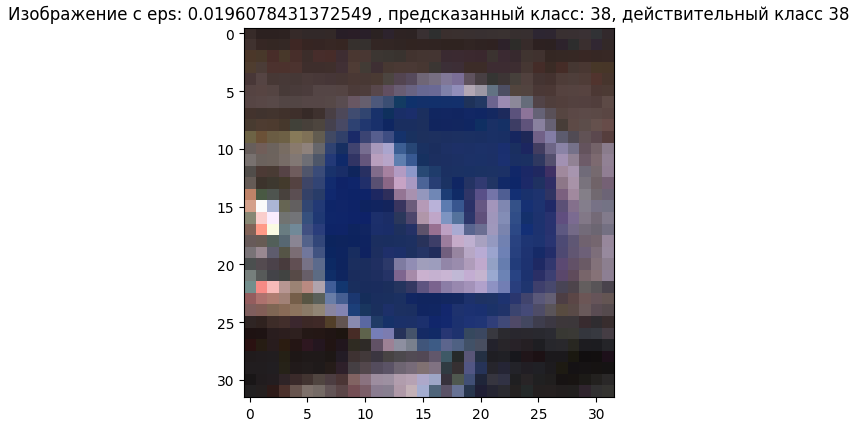


Рисунок 21 – Искаженное изображение VGG16

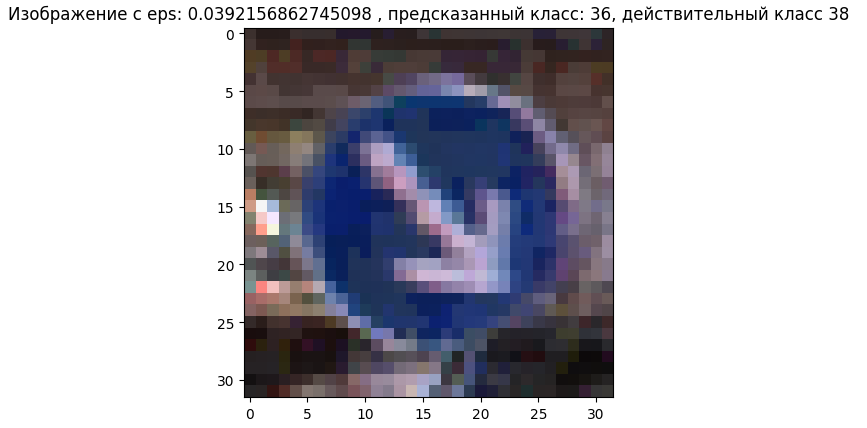


Рисунок 22 – Искаженное изображение VGG16

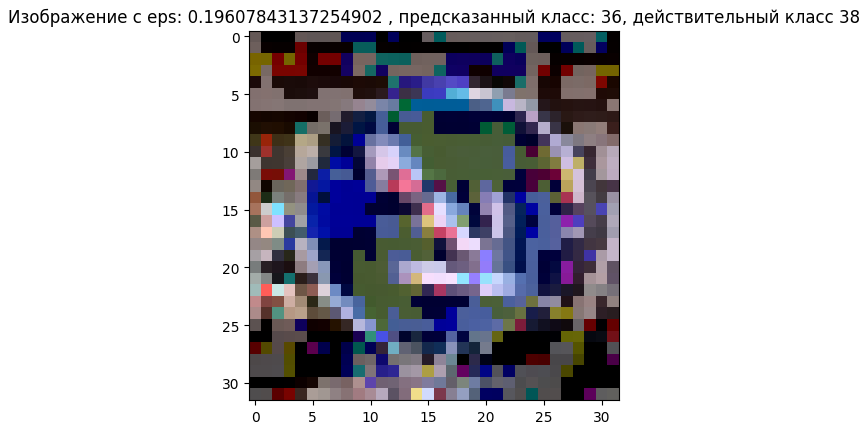


Рисунок 23 – Искаженное изображение VGG16

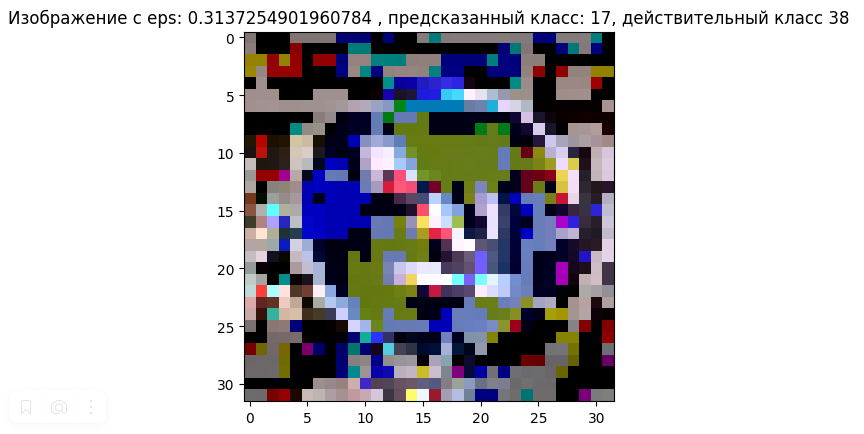


Рисунок 24 – Искаженное изображение VGG16

Таблица с результатами по заданию представлена ниже.

Таблица 2 – Таблица по результатам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Исходные изображения | Adversarial images 𝜖=1/255 | Adversarial images 𝜖=5/255 | Adversarial images 𝜖=10/255 |
| ResNet50 – FGSM | 91% | 74% | 33% | 17% |
| ResNet50 – PGD | 91% | 71% | 30% | 23% |
| VGG16 – FGSM | 89% | 79% | 44% | 21% |
| VGG16 – PGD | 89% | 77% | 48% | 32% |

# Задание 3

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

В данном задании требуется применить целевую атаку уклонения на основе белого ящика против моделей глубокого обучения.

Используем изображения знака «Стоп» (label class 14) из тестового набора данных. Всего имеется 270 изображений. Применим атаку Projected Gradient Descent (PGD) на знак «Стоп» с целью классификации его как знака «Ограничение скорости 30» (target label class = 1). Переберем значения искажений 𝜖 = [1/255, 3/255, 5/255, 10/255, 20/255, 50/255, 80/255].

Повторим атаку методом FGSM. Выведем 5 изображений классификации класса 14 как класс 1 при помощи целевой FGSM атаки. Изображения показаны на рисунках 25-33.



Рисунок 25 – Исходное изображение FGSM



Рисунок 26 – Искаженное изображение FGSM



Рисунок 27 – Исходное изображение FGSM

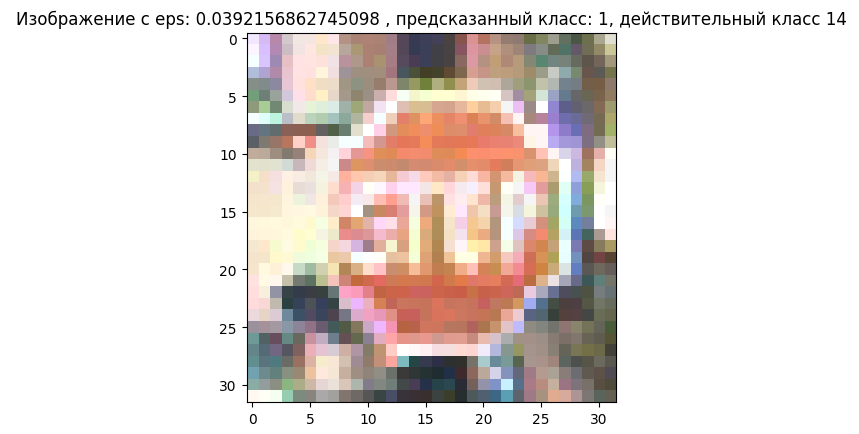


Рисунок 28 – Искаженное изображение FGSM



Рисунок 29 – Исходное изображение FGSM

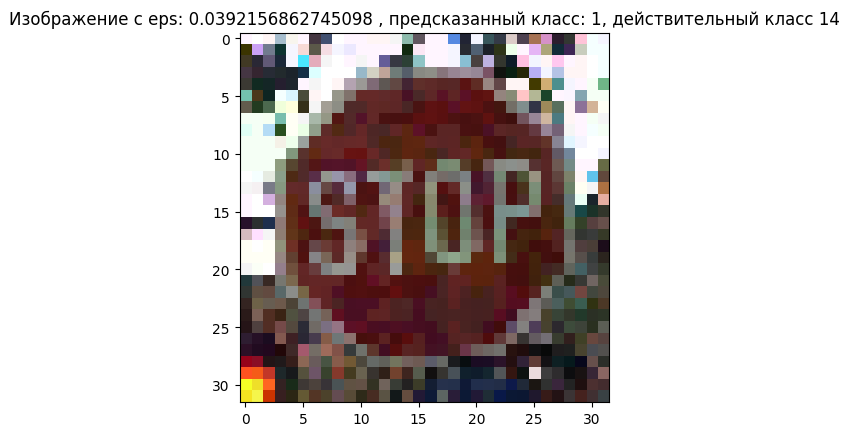


Рисунок 30 – Искаженное изображение FGSM



Рисунок 30 – Исходное изображение FGSM

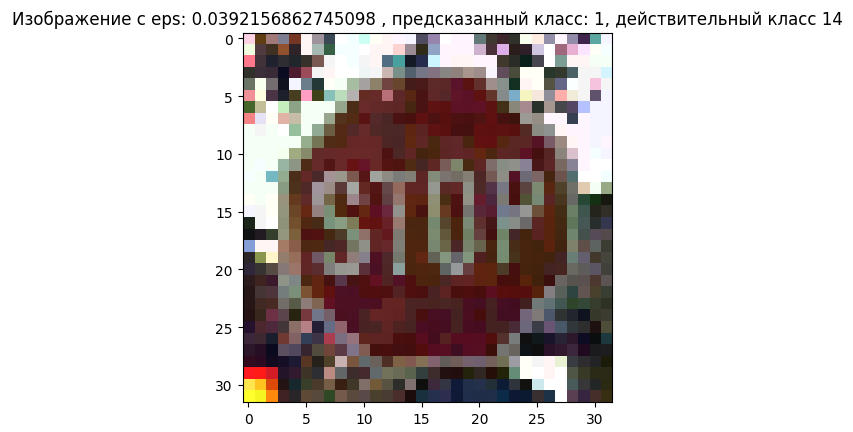


Рисунок 31 – Искаженное изображение FGSM



Рисунок 32 – Исходное изображение FGSM



Рисунок 33 – Искаженное изображение FGSM

Также повторим атаку методом PGD.. Выведем 5 изображений классификации класса 14 как класс 1 при помощи целевой PGD атаки. Изображения показаны на рисунках 34-43.



Рисунок 34 – Исходное изображение PGD

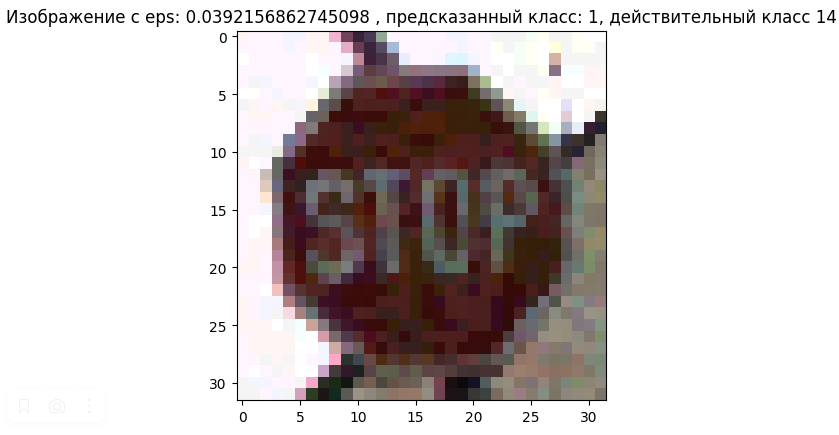


Рисунок 35 – Искаженное изображение PGD



Рисунок 36 – Исходное изображение PGD



Рисунок 37 – Искаженное изображение PGD



Рисунок 38 – Исходное изображение PGD



Рисунок 39 – Искаженное изображение PGD



Рисунок 40 – Исходное изображение PGD



Рисунок 41 – Искаженное изображение PGD



Рисунок 42 – Исходное изображение PGD

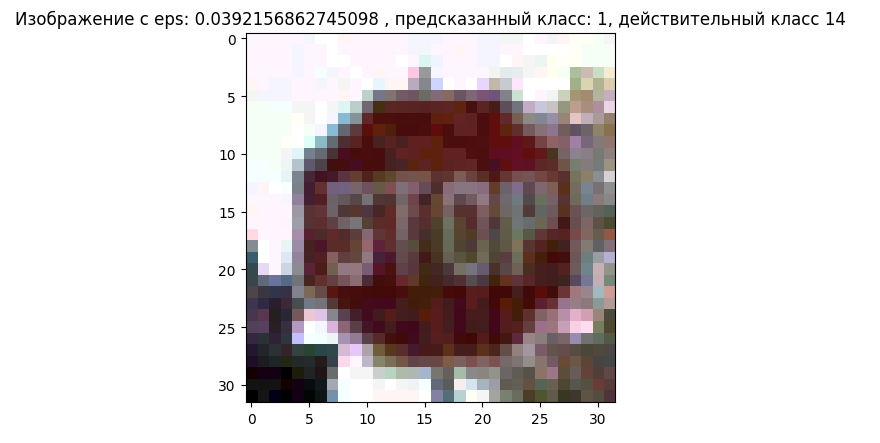


Рисунок 43 – Искаженное изображение PGD

После проделанных атак заполним таблицу точностей целевых атак FGSM и PGD на знак стоп. Результаты показаны в таблице 3.

Таблица 3 – Таблица по результатам

|  |  |  |
| --- | --- | --- |
| Искажение | PGD attack – Stop sign images | FGSM attack – Stop sign images |
| 𝜖=1/255 | 99% | 100% |
| 𝜖=3/255 | 90% | 85 |
| 𝜖=5/255 | 92% | 75 |
| 𝜖=10/255 | 70% | 27 |
| 𝜖=20/255 | 60% | 3 |
| 𝜖=50/255 | 1% | 0 |
| 𝜖=80/255 | 1% | 0 |

# Вывод

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 50/255.

# Заключение

В процессе выполнения лабораторной работы были реализованы указанные задачи, а именно:

- Созданы два классификатора с использованием глубоких нейронных сетей на наборе данных GTSRB;

- Применена нецелевая атака уклонения с использованием метода белого ящика на модели глубокого обучения;

- Применена целевая атака уклонения методом белого ящика на модели глубокого обучения.