|  |
| --- |
|  |
| 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-02-22

Исаев Александр Михайлович

**Проверил**:

К.т.н. Спирин Андрей Андреевич

Москва 2023

**Задание 1**

В рамках первого задания были обучены 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB.

Прежде всего, необходимо извлечь изображение для создания тренировочной выборки, процесс представлен на рисунке 1

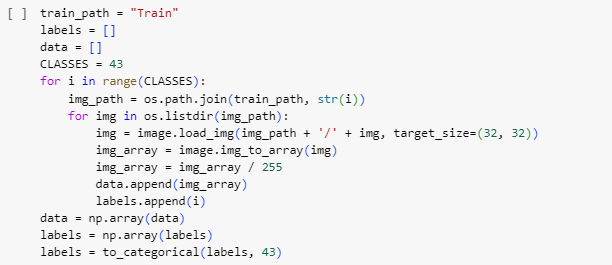


Рисунок 1 – Извлечение изображения

После чего отобразим изображение, отображение предствлено на рисунке 2

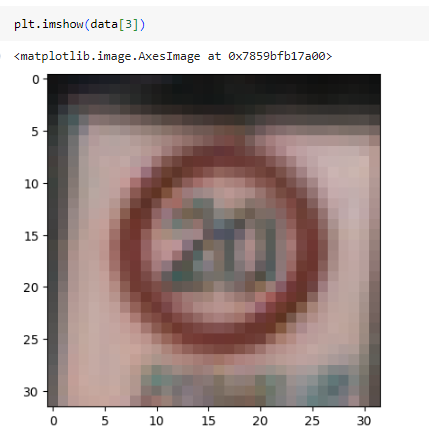


Рисунок 2 – Отображение изображения

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 3

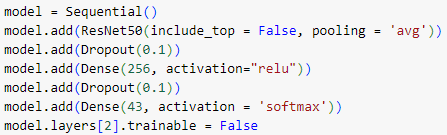


Рисунок 3 – Модель ResNet50

Графики точности, график потерь представлены на рисунках 4 и 5. Валидационные показатели ResNet50 представлены на рисунке 6.

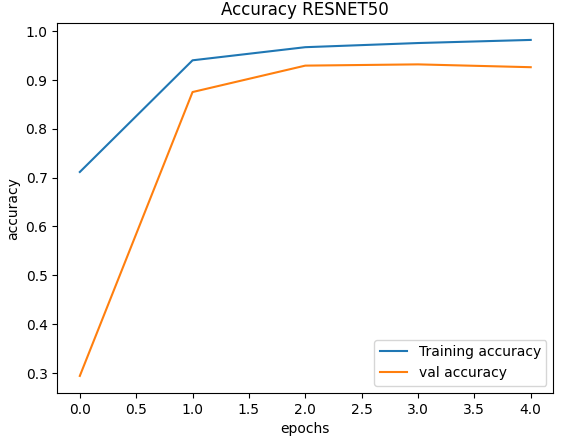


Рисунок 4 – График точности

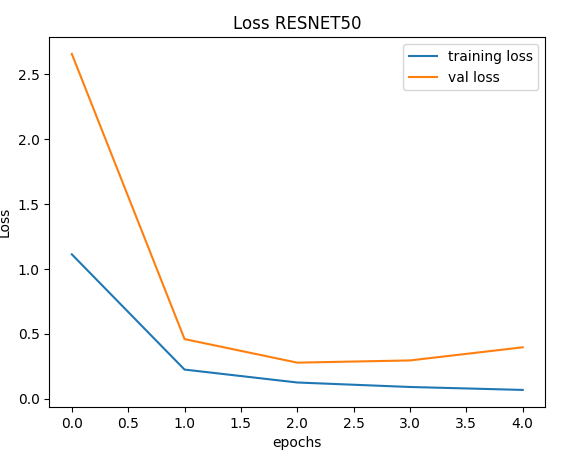


Рисунок 5 – График потерь



Рисунок 6 – Валидация ResNet50

После обучения модель была протестирована на тестовом наборе. Показатели валидации приведены на рисунке 7.



Рисунок 7 – Тестирование ResNet50

Вторая модель построена на базе VGG16 и состоит из слоев, представленных на рисунке 8.

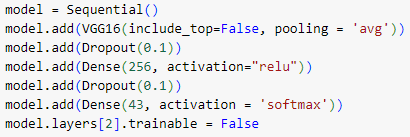


Рисунок 8 – Модель VGG16

Графики процесса обучения: график точности и график потерь представлены на рисунках 9 и 10.

Валидационные показатели VGG16 представлены на рисунке 11.

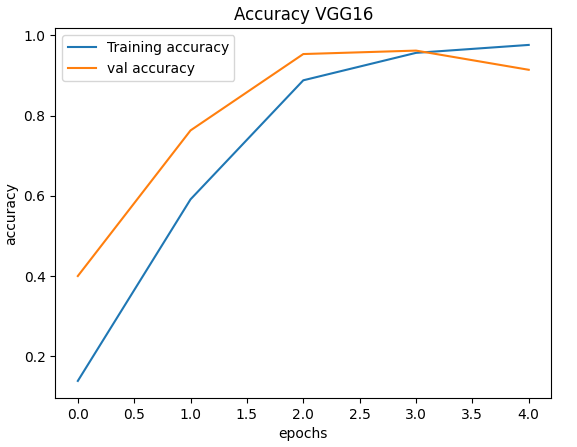


Рисунок 9 – Графики точности

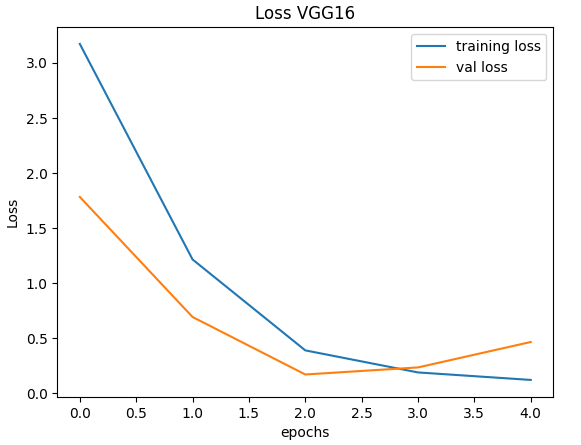


Рисунок 10 – График потерь



Рисунок 11 – Валидация VGG16

После обучения модель была протестирована на валидационном наборе. Показатели валидации приведены на рисунке 12.

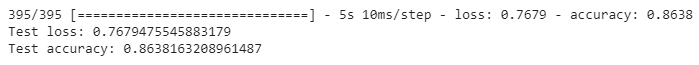


Рисунок 12 – Тестировние VGG16

Итоговая таблица по заданию представлена под номером 1.

Таблица 1 – Таблица по результатам

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Обучение** | **Валидация** | **Тест** |
| ResNet50 | loss: 0.0677 accuracy: 0.9818 | loss: 0.3958 accuracy: 0.9260 | loss: 0.7337 accuracy: 0.8681 |
| VGG16 | loss: 0.1230 accuracy: 0.9762 | loss: 0.4669 accuracy: 0.9141 | loss: 0.7679 accuracy: 0.8638 |

# **Задание 2**

В рамках задания предстоит 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 13.

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] (рисунки 14-16).

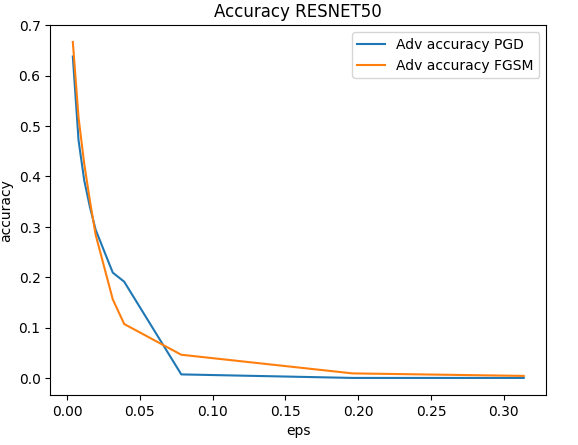


Рисунок 13 – График ResNet50

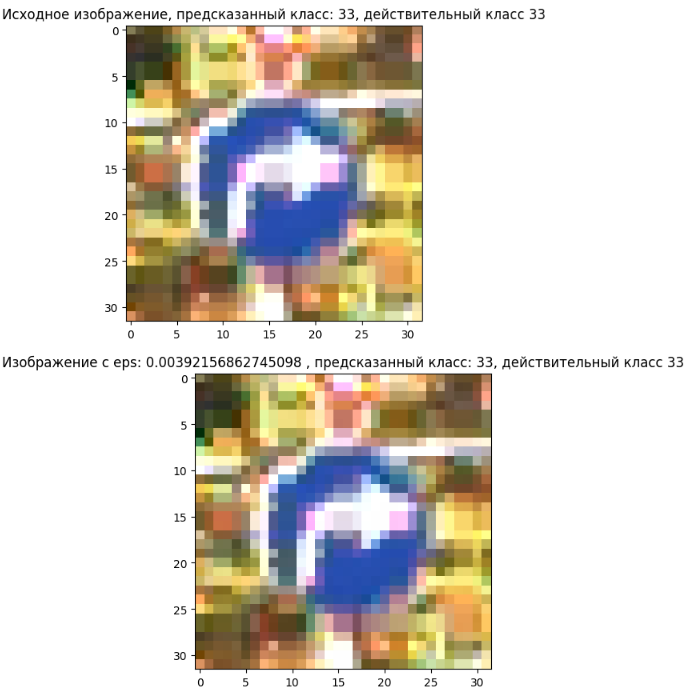


Рисунок 14– Исходное и искаженные изображения ResNet50

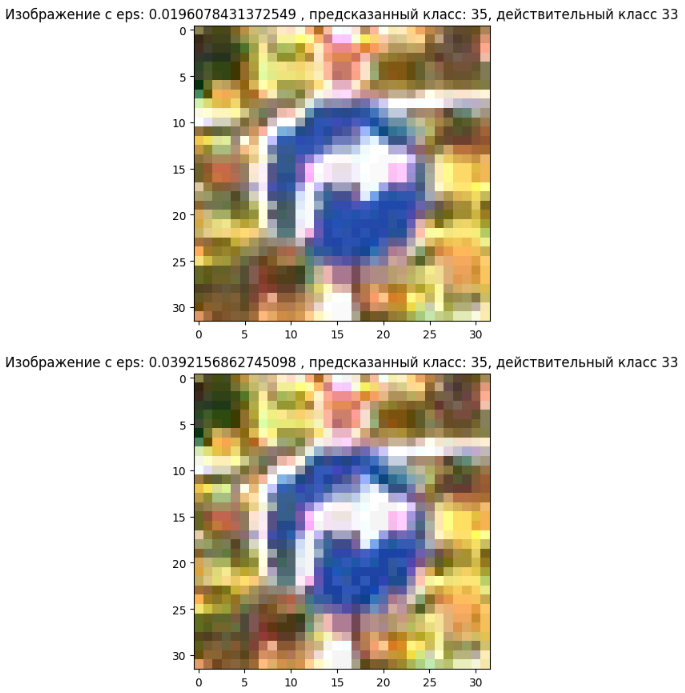


Рисунок 15– Искаженные изображения ResNet50

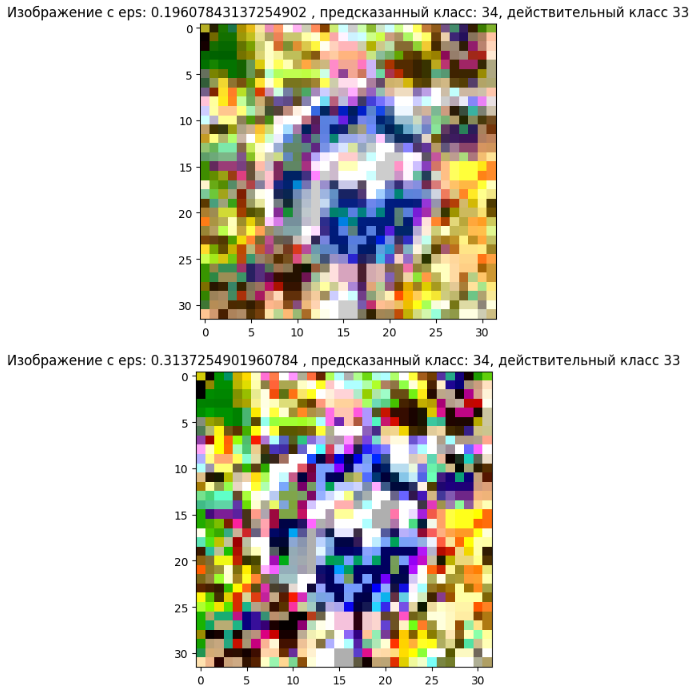


Рисунок 16 – Искаженные изображения ResNet50

Проделаем повторный 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 17.

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] (рисунки 2.6-2.8).

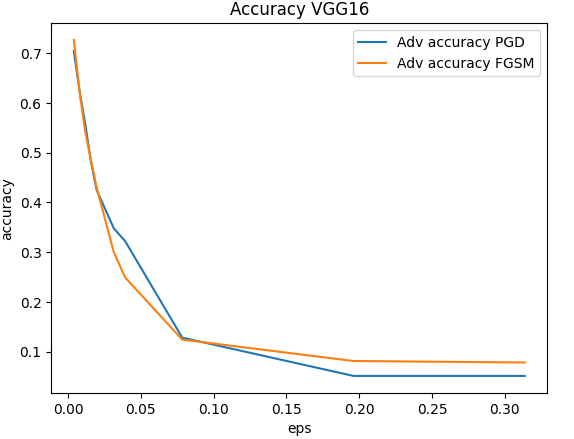


Рисунок 17 – График VGG16

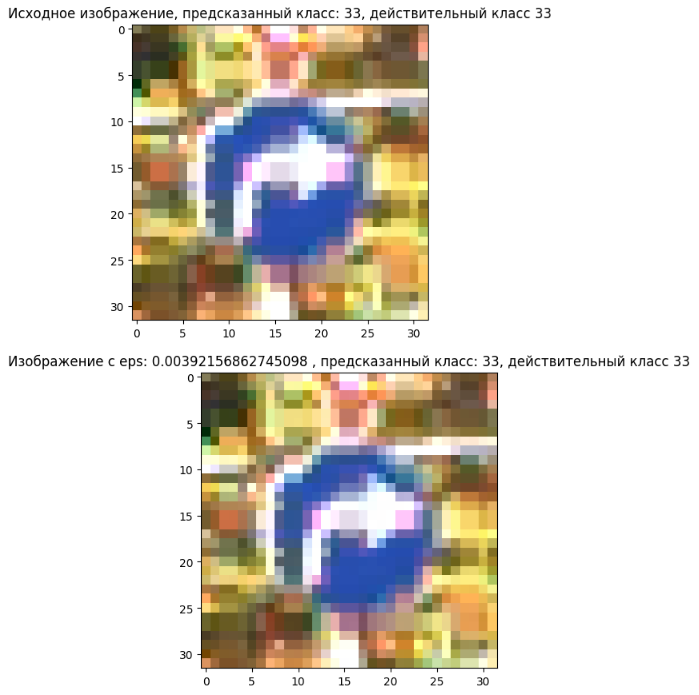


Рисунок 18 – Исходное и искаженные изображения VGG16

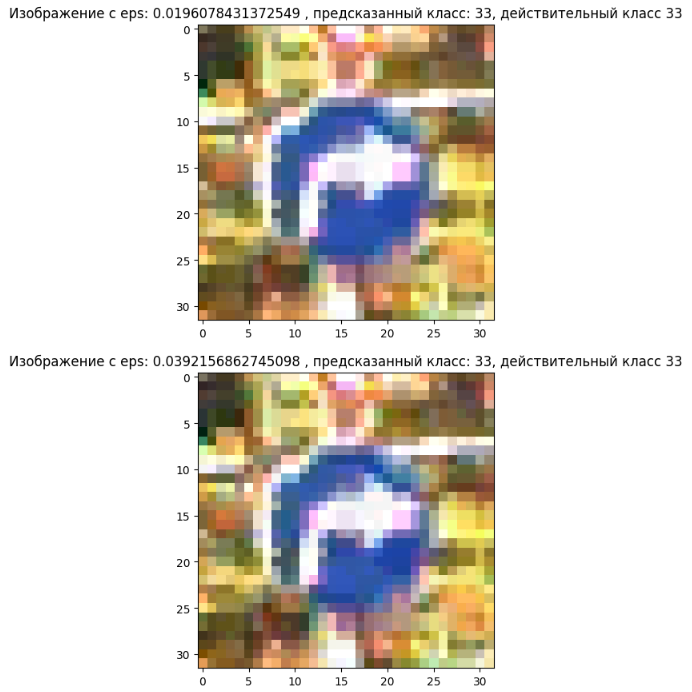


Рисунок 19 – Искаженные изображения VGG16

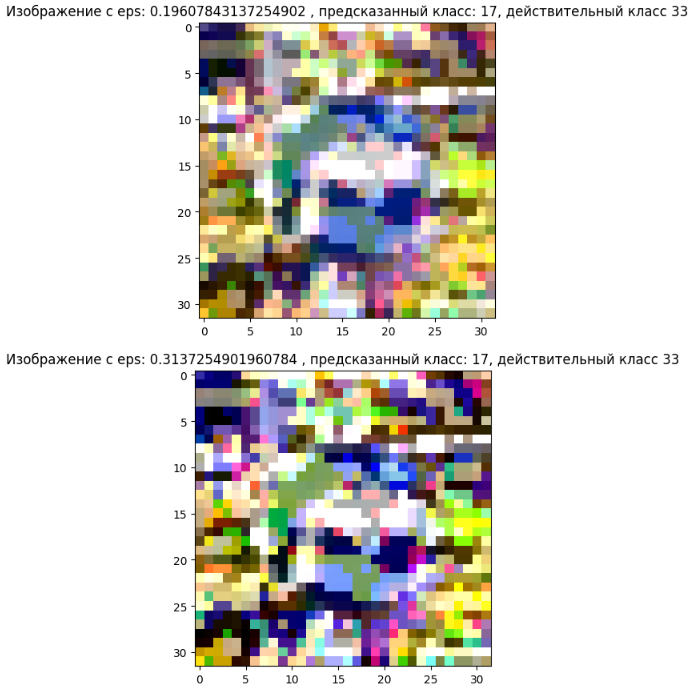


Рисунок 20 – Искаженные изображения VGG16

Результирующая таблица по заданию представлена под номером 2

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 | loss: 0.0677  accuracy: 0.9818 | loss: 2.5200  accuracy: 0.6669 | loss: 7.5104  accuracy: 0.2840 | loss: 9.3754  accuracy: 0.1070 |
| ResNet50 – PGD | loss: 0.0677  accuracy: 0.9818 | loss: 2.8430  accuracy: 0.6380 | loss: 7.8228  accuracy: 0.2939 | loss: 9.7428  accuracy: 0.1909 |
| VGG16 – FGSM | loss: 0.1230  accuracy: 0.9762 | loss: 2.0048  accuracy: 0.7269 | loss: 5.6454  accuracy: 0.4300 | loss: 7.6370  accuracy: 0.2489 |
| VGG16 – PGD | loss: 0.1230  accuracy: 0.9762 | loss: 2.4107  accuracy: 0.7039 | loss: 8.1499  accuracy: 0.4250 | loss: 10.8968  accuracy: 0.3219 |

# **Задание 3**

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], и заполним таблицу 3 значениями точности классификации изображений знаков "Стоп" и "Ограничение скорости 30".

Повторим атаку методом FGSM, и заполним таблицу 3

Таблица 3 – Таблица по результатам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Искажение** | **PGD attack – Stop sign images** | **PGD attack – Speed Limit 30 sign images** | **FGSM attack – Stop sign images** | **FGSM attack – Speed Limit 30 sign images** |
| 𝜖=1/255 | loss: 0.1990  accuracy: 0.9555 | loss: 0.0604  accuracy: 0.9777 | loss: 0.6644  accuracy: 0.8592 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=3/255 | loss: 0.9776  accuracy: 0.8444 | loss: 0.0604  accuracy: 0.9777 | loss: 2.2754  accuracy: 0.6370 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=5/255 | loss: 1.8796  accuracy: 0.6925 | loss: 0.0604  accuracy: 0.9777 | loss: 3.7947  accuracy: 0.4111 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=10/255 | loss: 3.0420  accuracy: 0.5185 | loss: 0.0604  accuracy: 0.9777 | loss: 6.6825  accuracy: 0.0703 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=20/255 | loss: 7.5241  accuracy: 0.2481 | loss: 0.0604  accuracy: 0.9777 | loss: 7.7221  accuracy: 0.0074 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=50/255 | loss: 14.7153  accuracy: 0.0555 | loss: 0.0604  accuracy: 0.9777 | loss: 7.1935  accuracy: 0.0074 | loss: 0.0604  accuracy: 0.9777 |
| 𝜖=80/255 | loss: 15.4812  accuracy: 0.0259 | loss: 0.0604  accuracy: 0.9777 | loss: 7.2758  accuracy: 0.0074 | loss: 0.0604  accuracy: 0.9777 |

Метод FGSM плохо подходит для целевых атак. С ростом искажения классификация начинает давать сбои. Оптимальным значением искажения является 10/255. При больших значения модель будет ошибаться всегда.

PGD отлично подходит для целевых атак. При больших искажениях, модель почти всегда будет определять заданный нами класс, но изображение станет слишком навязчиво искажено. Оптимальным значением искажения является 20/255.

# Заключение

Таким образом, в ходе выполнения лабораторной работы №2 были созданы 2 классификатора на основе глубоких нейронных сетей на основе датасета GTSRB, выявлена и отображена их точность и графики точности, потерь и валидации.

Была применена нецелевая атака уклонения на основе белого ящика против моделей глубокого обучения: были отображены искаженные изображения. Была составлена таблица по итогом задания 2.

Была применена целевая атака уклонения на основе белого ящика против моделей глубокого обучения и отображена финальная таблица по результатам атак.