

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

**«МИРЭА – Российский технологический университет»** **РТУ МИРЭА**

Кафедра: КБ-4 «Киберразведка и противодействие угрозам с применением технологий искусственного интеллекта»

**Лабораторная работа №2 по дисциплине**

**«Анализ защищенности систем искусственного интеллекта»**

Выполнил Морин А.А.

Группа:

ББМО-01-23

Москва, 2024 г.

**Задание**

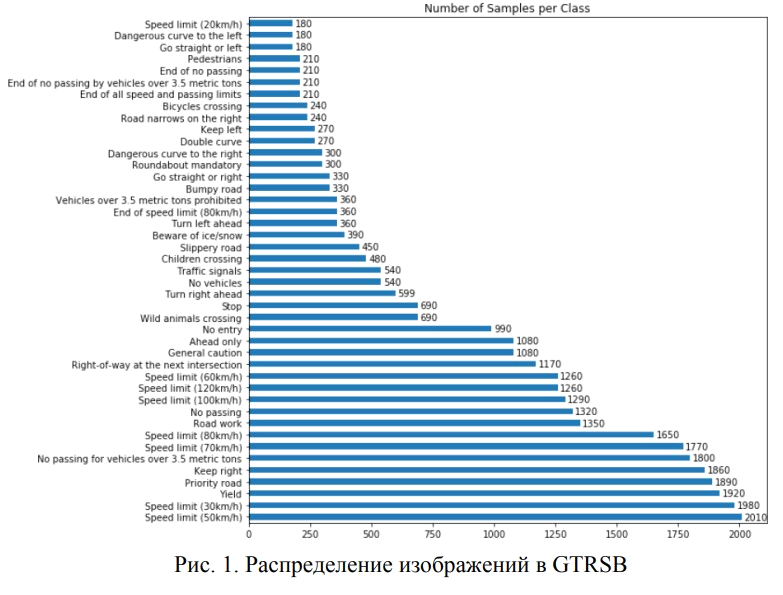
Задачи:

* 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. Распределение изображений по классам показано на рис. 1. Набор данных:

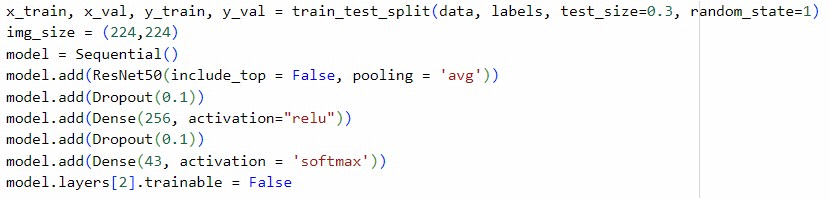
<https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german>-

traffic-sign

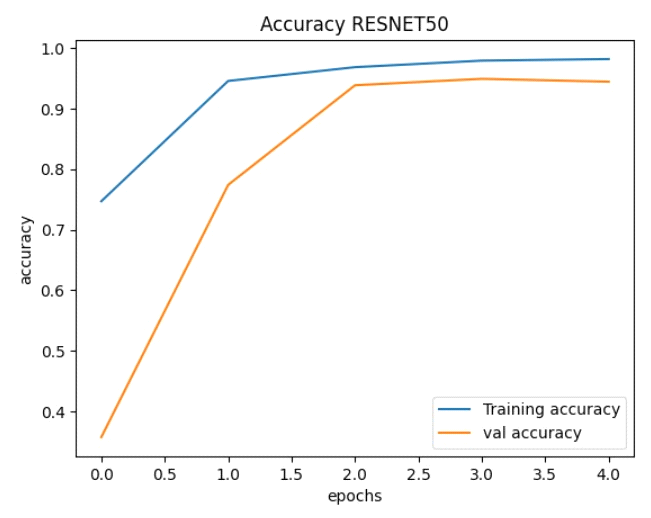


**Задание 1**

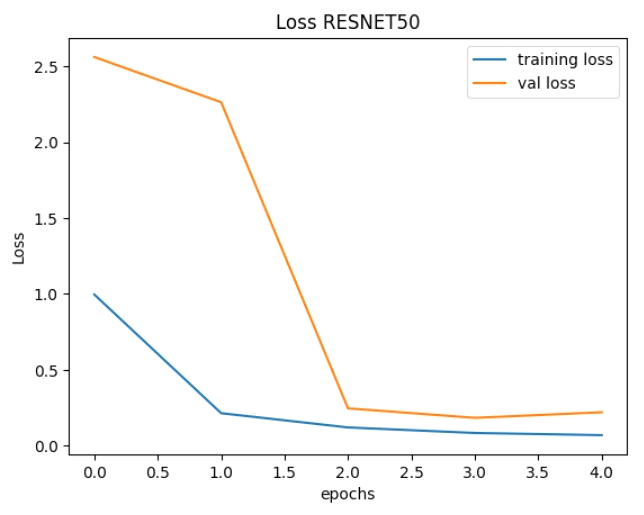
Создаем модель ResNet50, выборки поделены 70/30



Первый график отображает точность обучения и валидации модели RESNET50



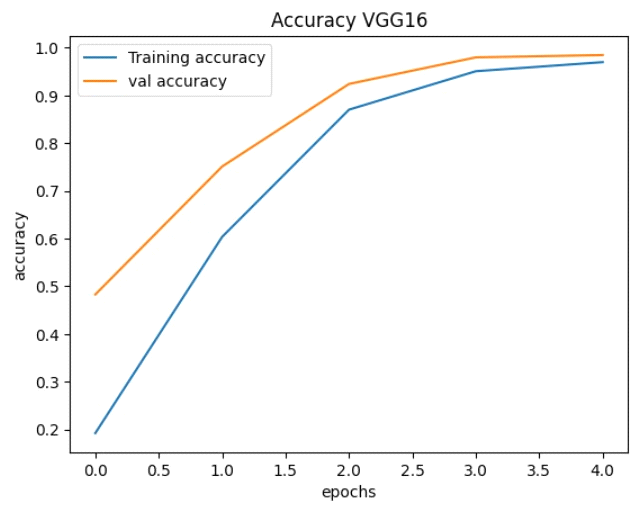
Второй график отображает потерю обучения и валидации модели RESNET50



Создаем модель VGG16



Третий график отображает точность обучения и валидации модели VGG16



Четвертый график отображает потерю обучения и валидации модели VGG16



|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Обучение | Валидация | Тест |
| ResNet50 | loss: 0.0697 accuracy: 0.9816 | loss: 0.2205 accuracy: 0.9442 | loss: 0.4797 accuracy: 0.8907 |
| VGG16 | loss: 0.1551 accuracy: 0.9698 | loss: 0.0592 accuracy: 0.9847 | loss: 0.2825 accuracy: 0.9426 |

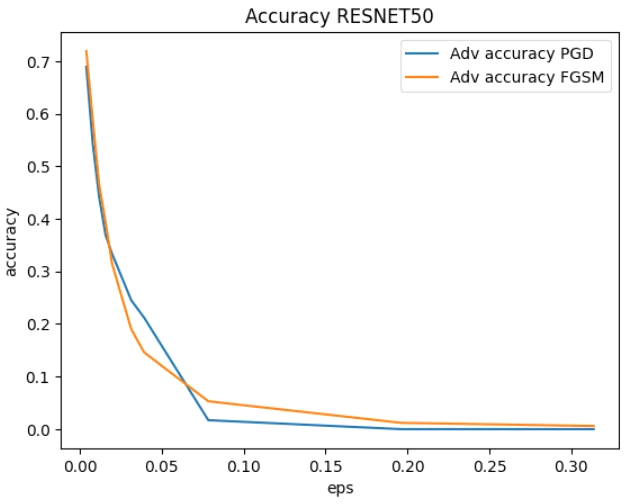
**Задание 2**

Проведем атаки FGSM и PGD на модель RESNET50, используя первые

1,000 изображений из тестового множества. Используем значения параметра искажения:

𝜖 = [1/255, 2/255, 3/255, 4/255, 5/255, 8/255, 10/255, 20/255, 50/255, 80/255].

Построим график зависимости точности классификации от параметра искажений эпсилон для RESNET50

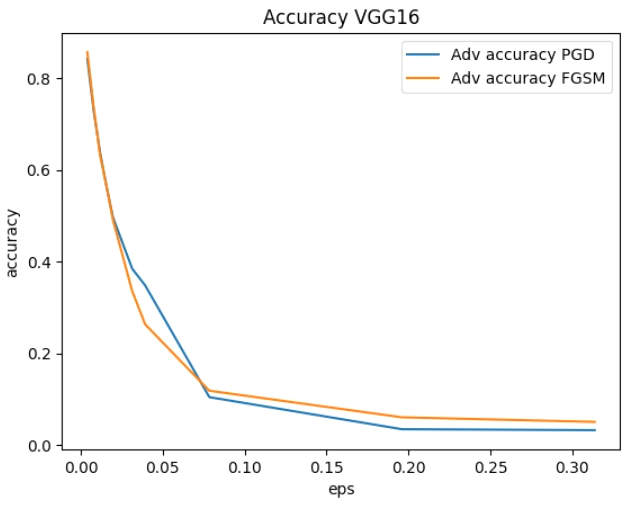


Проведем атаки FGSM и PGD на модель VGG16, используя первые

1,000 изображений из тестового множества. Используем значения параметра искажения:

𝜖 = [1/255, 2/255, 3/255, 4/255, 5/255, 8/255, 10/255, 20/255, 50/255, 80/255].

Построим график зависимости точности классификации от параметра искажений эпсилон для VGG16



Для атаки FGSM RESNET50, отобразим исходное изображение из датасета и атакующее изображение с указанием величины параметра = [1/255, 5/255, 10/255, 50/255, 80/255], также отобразим предсказанный класс атакующего изображения

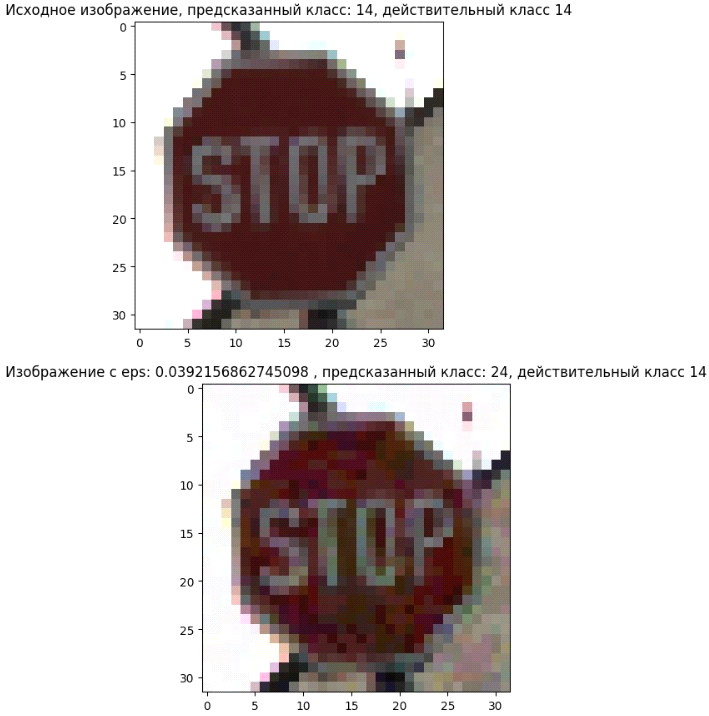
Для атаки FGSM VGG16, отобразим исходное изображение из датасета и атакующее изображение с указанием величины параметра = [1/255, 5/255, 10/255, 50/255, 80/255], также отобразим предсказанный класс атакующего изображения

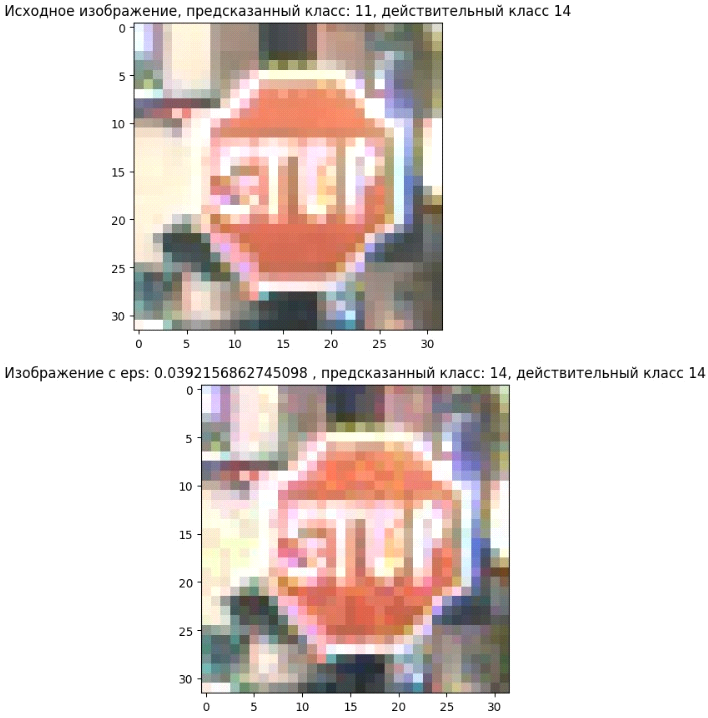
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Исходные изображения | Adversarial images  =1/255 | Adversarial images  =5/255 | Adversarial images  =10/255 |
| VGG16 -  FGSM | 89% | 79% | 44% | 21% |
| VGG16 -  PGD | 89% | 77% | 48% | 32% |
| ResNet50 -  FGSM | 91% | 74% | 33% | 17% |
|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

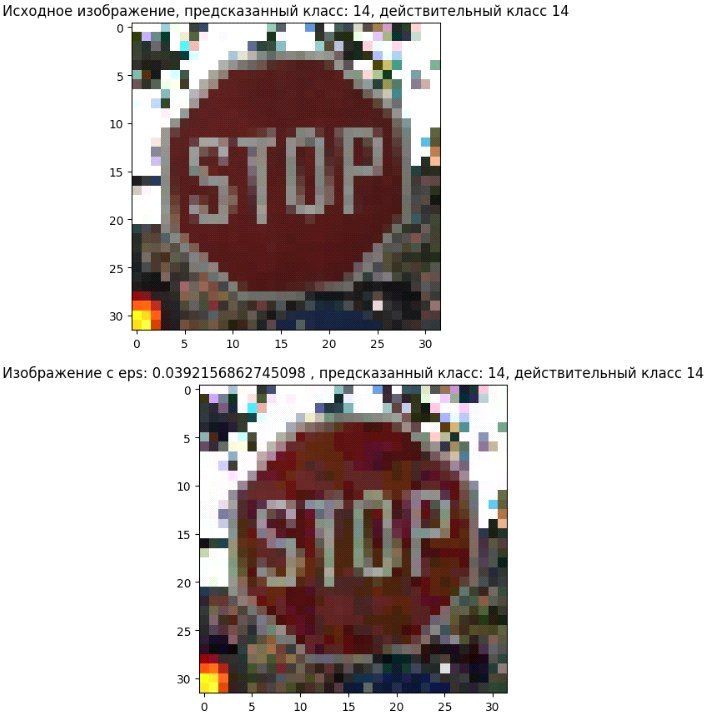
**Задание 3**

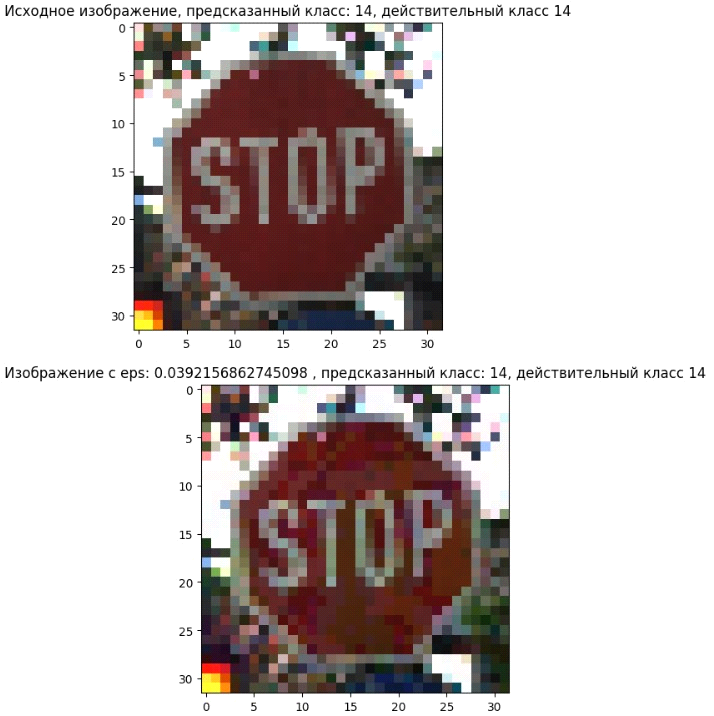
Используя изображения знака «Стоп» (label class 14) из тестового набора данных, применим атаки FGSM и PGD на знак «Стоп» с целью классификации его как знака «Ограничение скорости 30» (target label class = 1), изменяя значения искажений = [1/255, 3/255, 5/255, 10/255, 20/255, 50/255, 80/255].

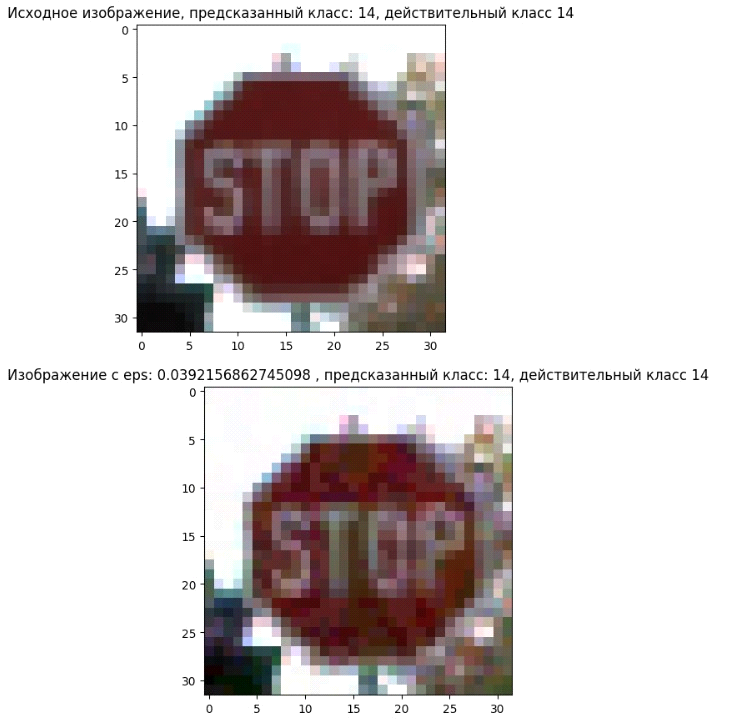
Выведем 5 пар примеров исходных изображений знака «Стоп» и соответствующих атакующих примеров для атаки FGSM



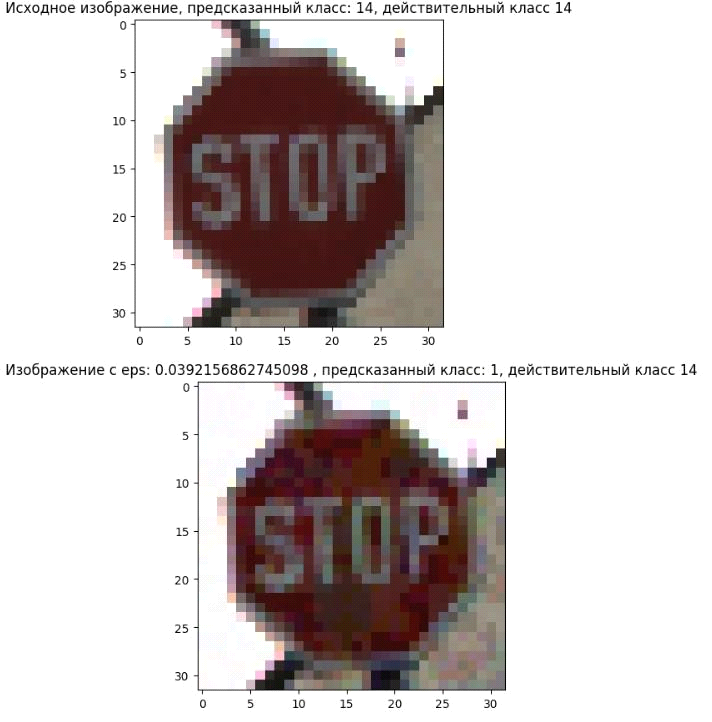


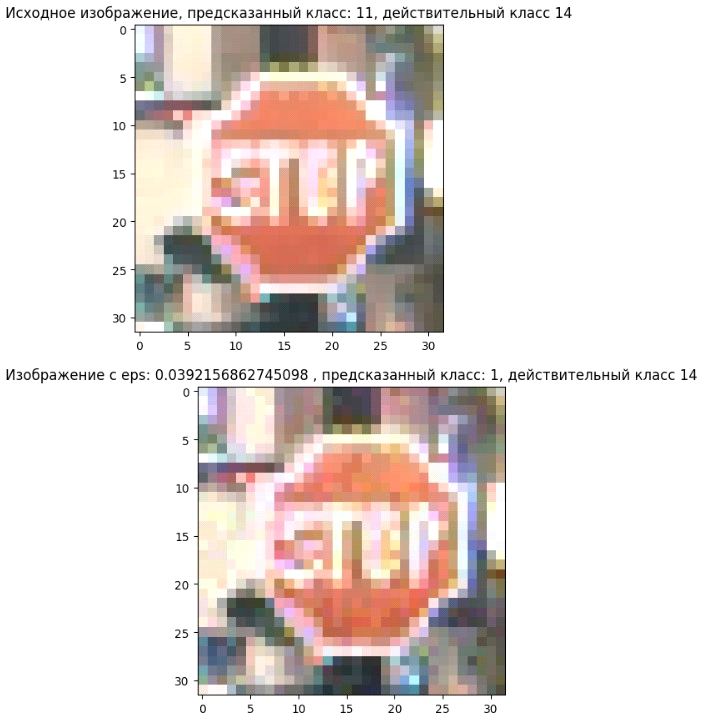


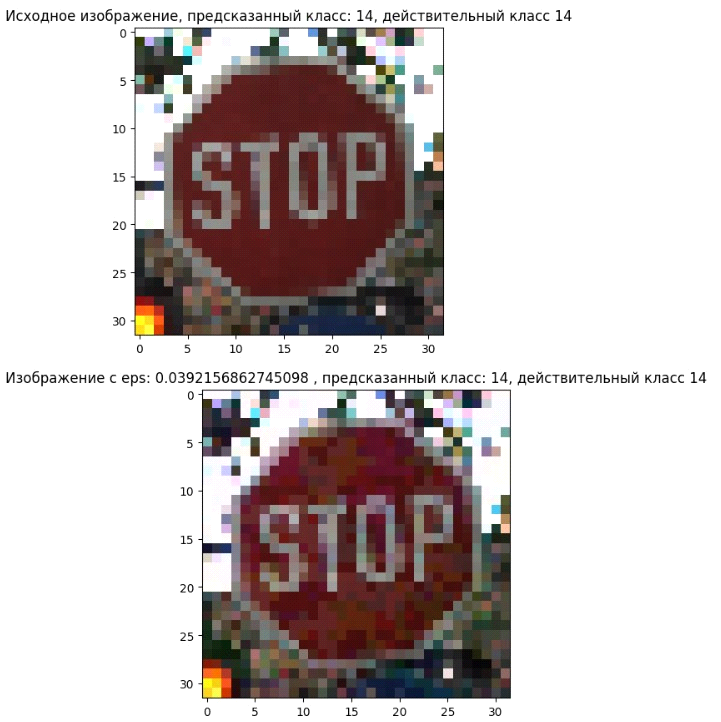


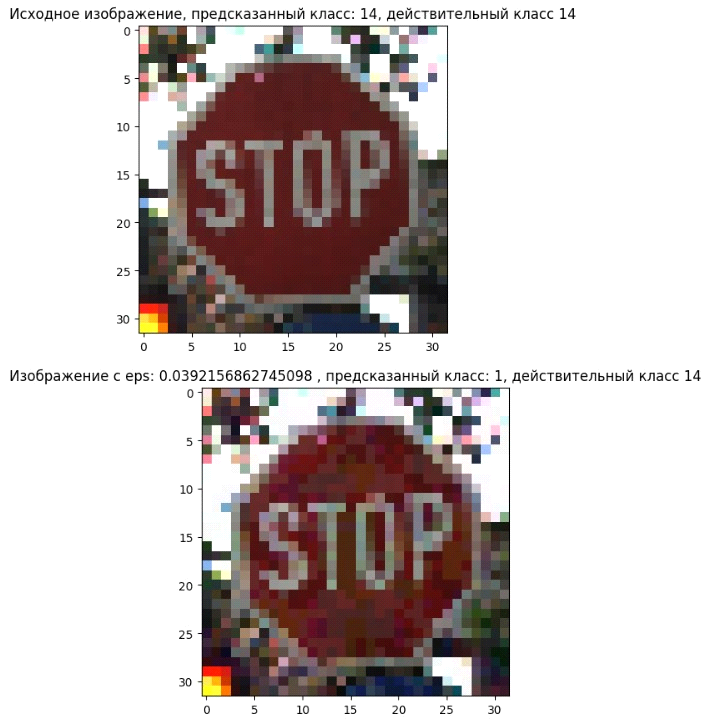


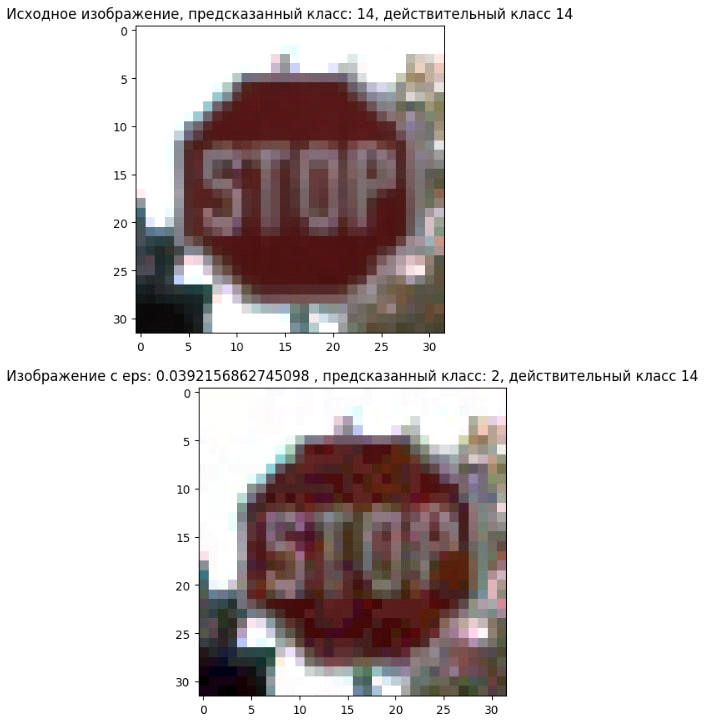
Выведем 5 пар примеров исходных изображений знака «Стоп» и соответствующих атакующих примеров для атаки PG











|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Искажение | FGSM – Stop | FGSM – Limit 30 | PGD - Stop | PGD – Limit 30 |
| 1/255 | 99% | 99% | 97% | 99% |
| 3/255 | 80% | 99% | 91% | 99% |
| 5/255 | 73% | 99% | 90% | 99% |
| 10/255 | 26% | 99% | 71% | 99% |

По результатам видно метод PGD значительно лучше подходит для целевой атаки, чем метод FGSM.

**Выводы**

В ходе работы были реализованы атаки уклонения на основе белого ящика против классификационных моделей на основе глубокого обучения и получены практические навыки переноса атак уклонения на основе черного ящика против моделей машинного обучения.

В целом, работа демонстрирует эффективность атак уклонения на основе белого ящика против моделей машинного обучения и необходимость дальнейших исследований в области безопасности систем ИИ.