|  |
| --- |
|  |
| 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

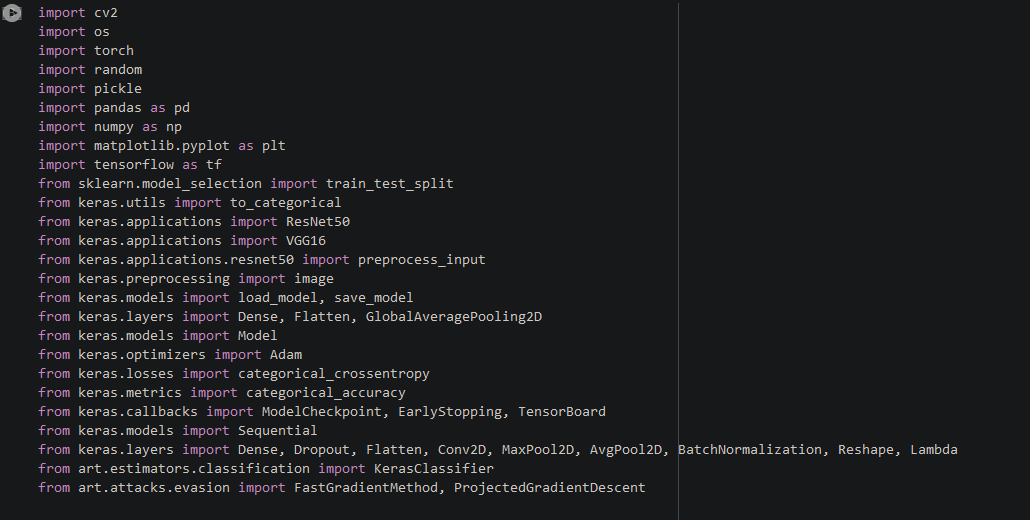
Простяков Никита Артемович

**Проверил**:

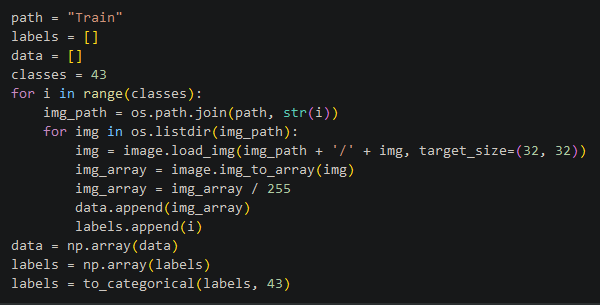
Спирин Андрей Андреевич

Москва 2023

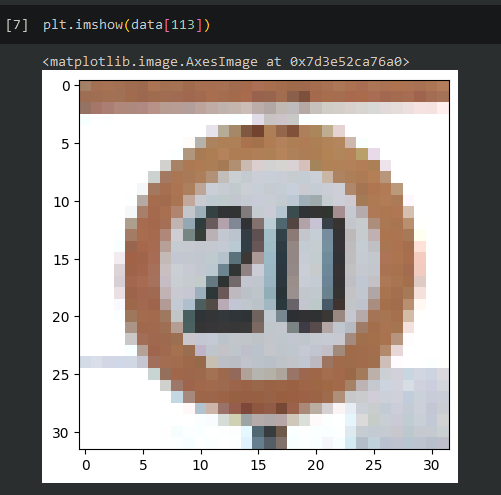
Импорт необходимых библиотек



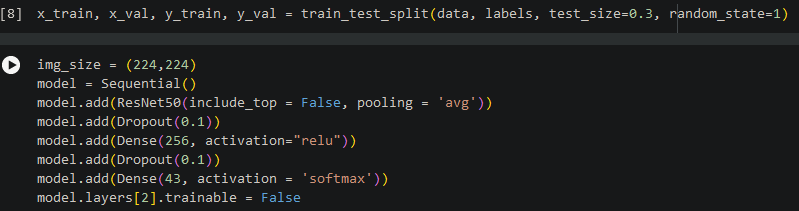
Загрузка изображений



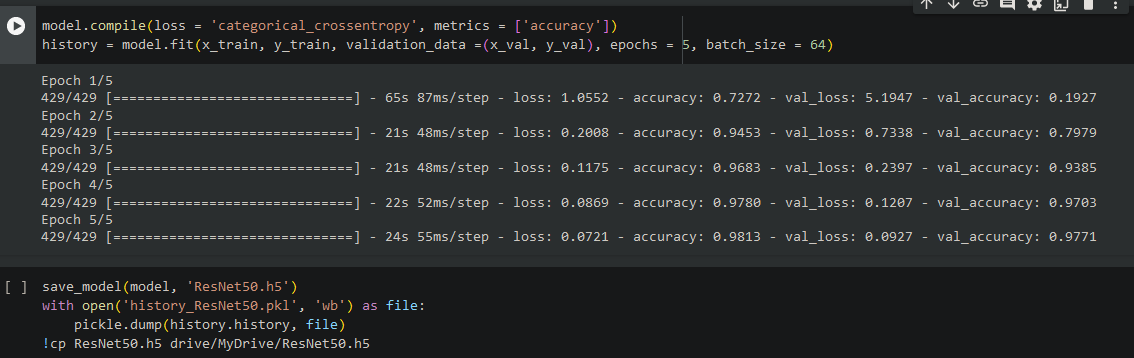
Пример изображения



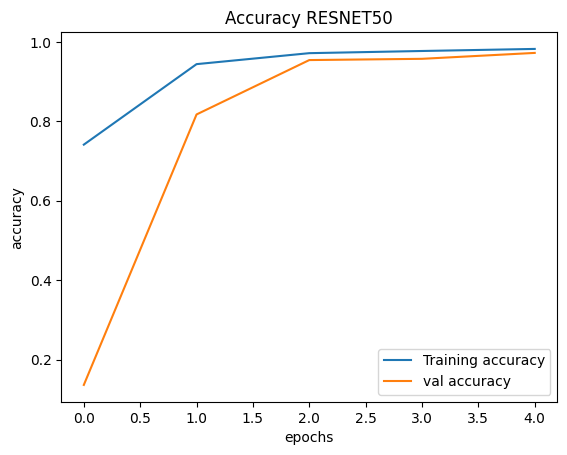
Разбиение на обучающую и тестовые выборки

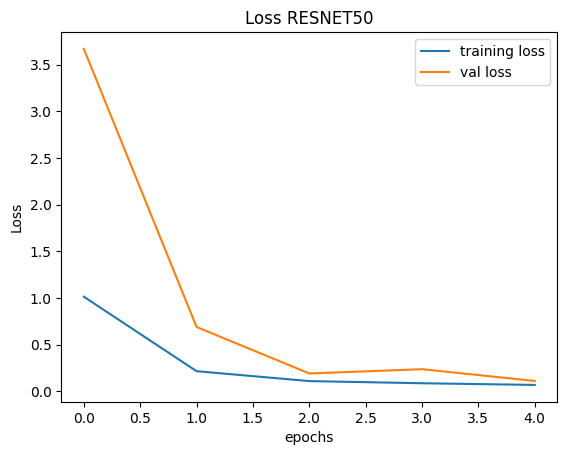


Обучение модели RESNET50

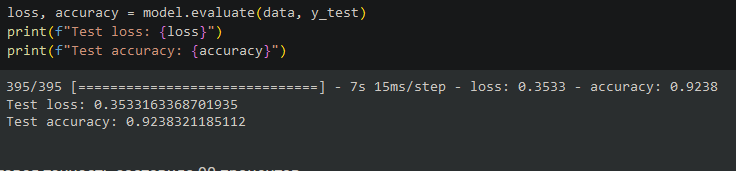


Графики точности и потерь модели в зависимости от эпохи

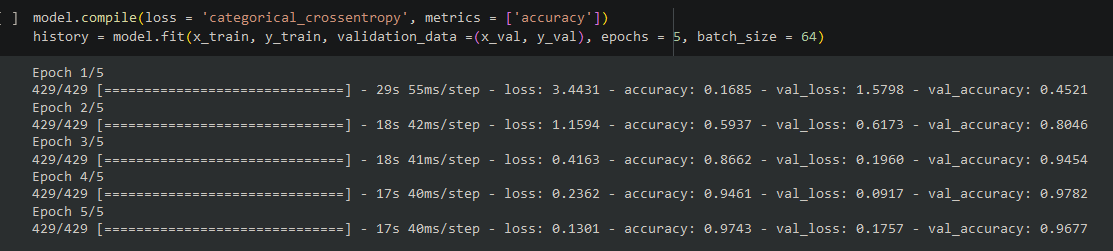




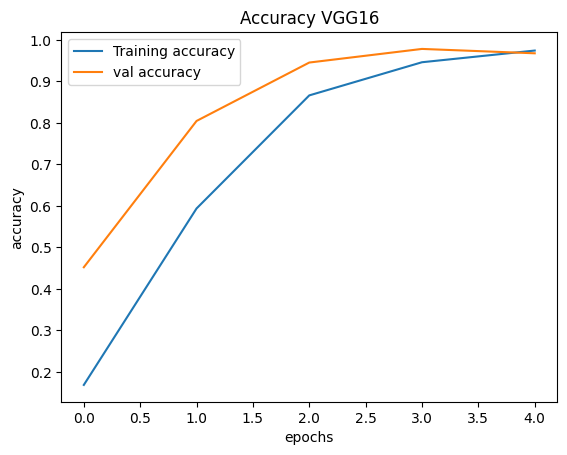
Итоговые точность и потери

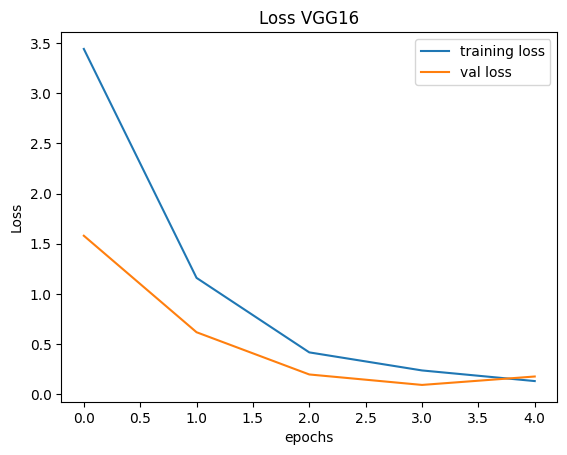


Обучение модели VGG16

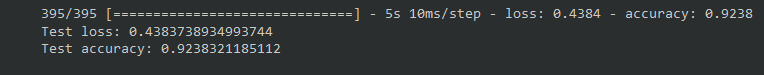


Графики точности и потерь VGG16

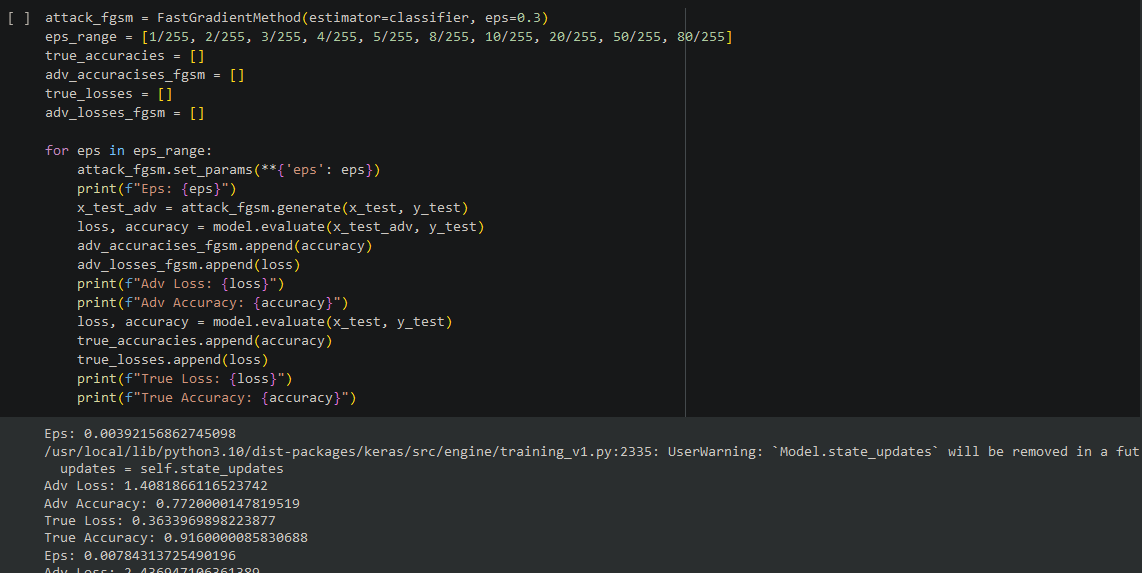




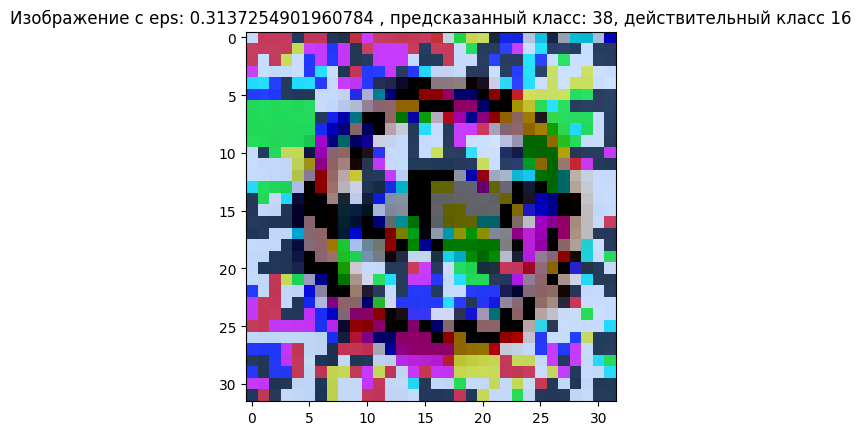
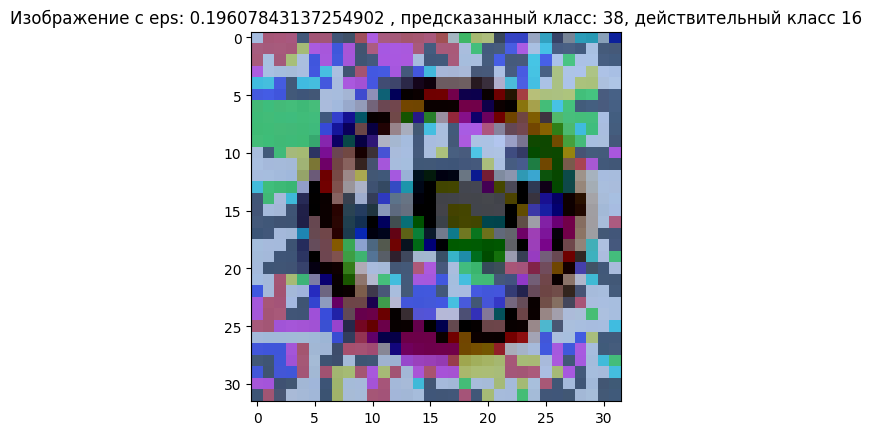
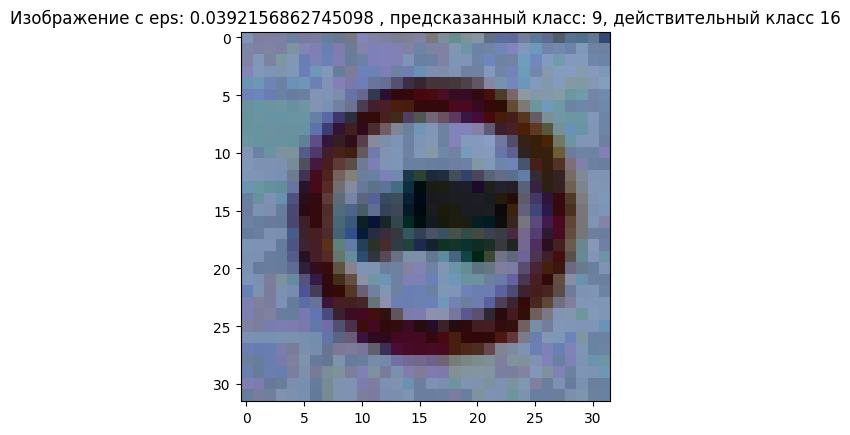
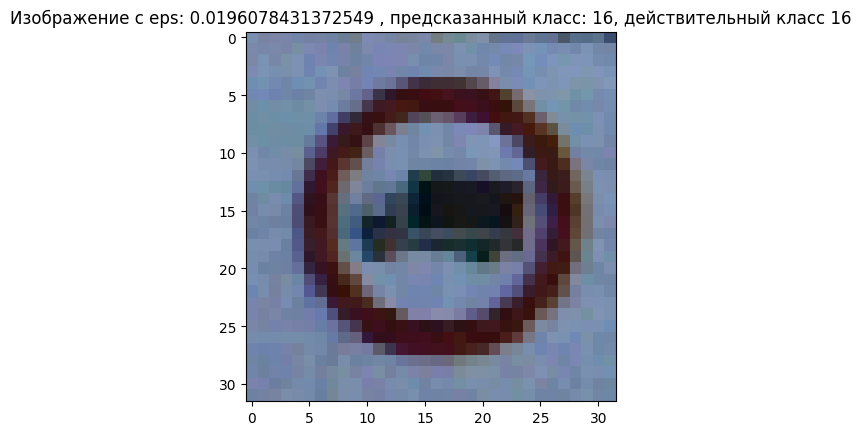
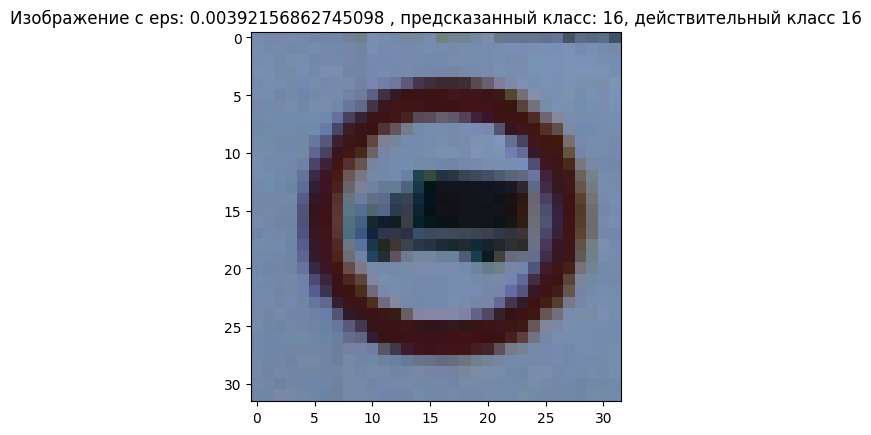
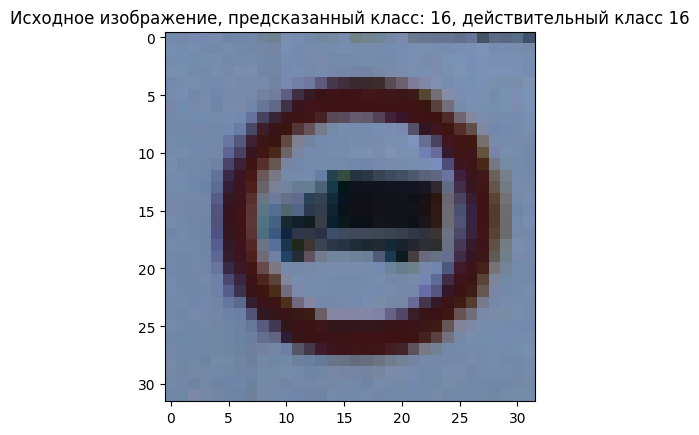
Итоговые точность и потери VGG16



Проведение атаки FGSM на RESNET50

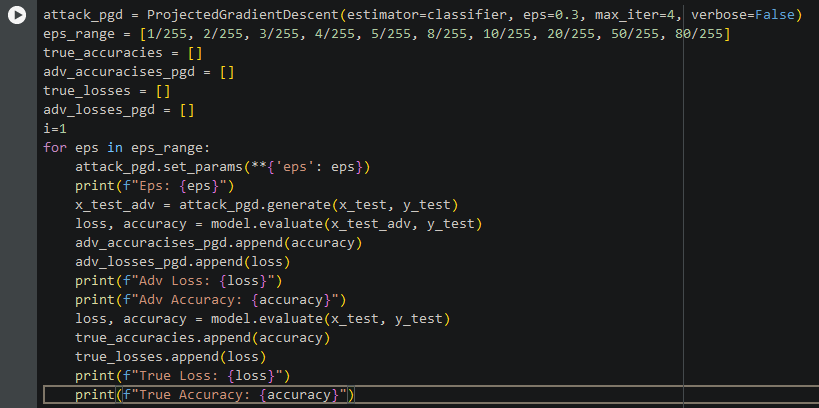


Изображения с различным значением eps

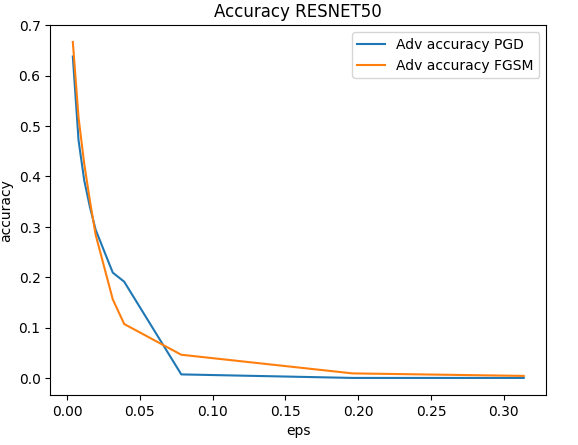


Оптимальным значением eps является 0.04, так как при данном значении наименее заметно вмешательство в изображение

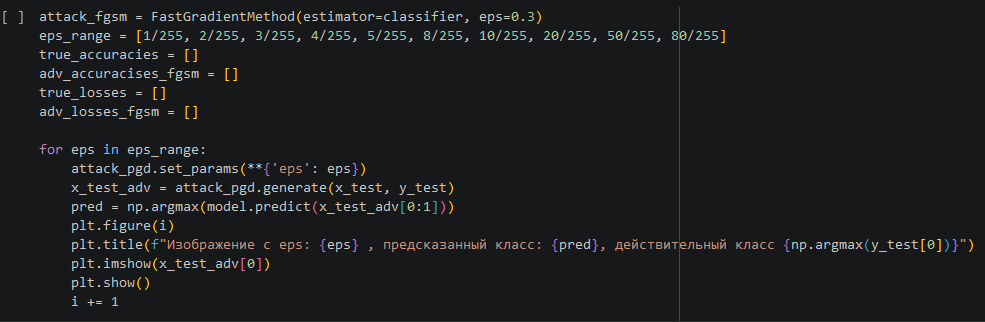
Атака PGD на RESNET50



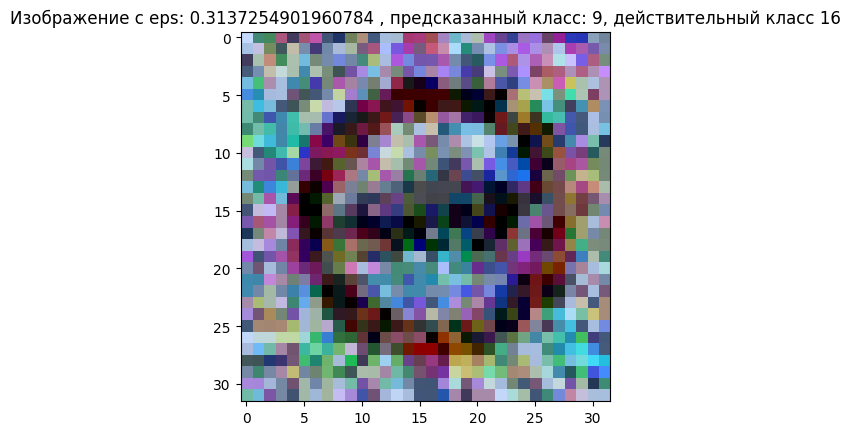
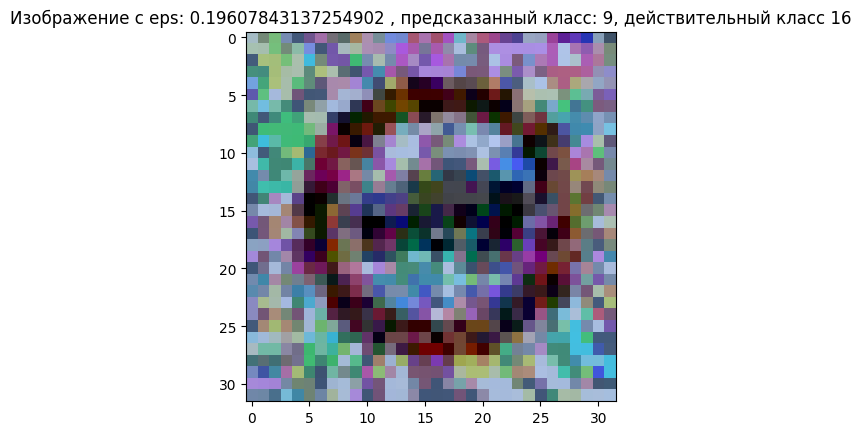
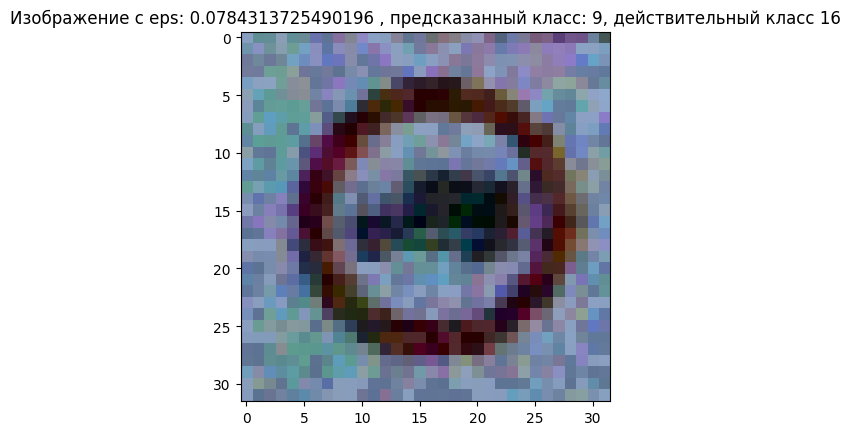
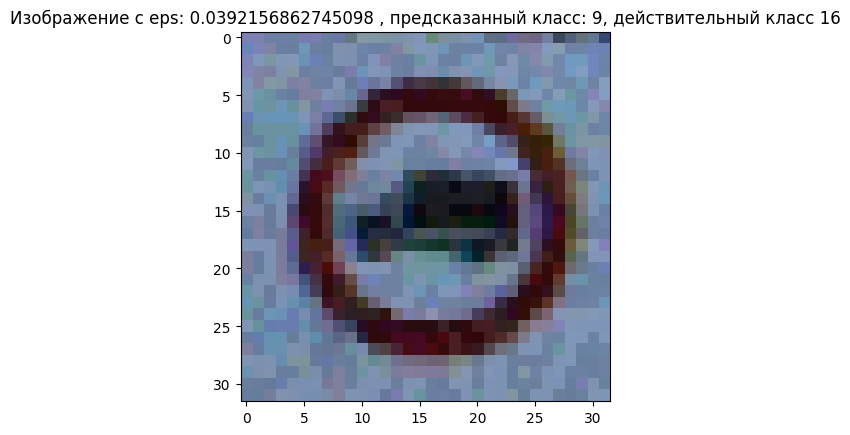
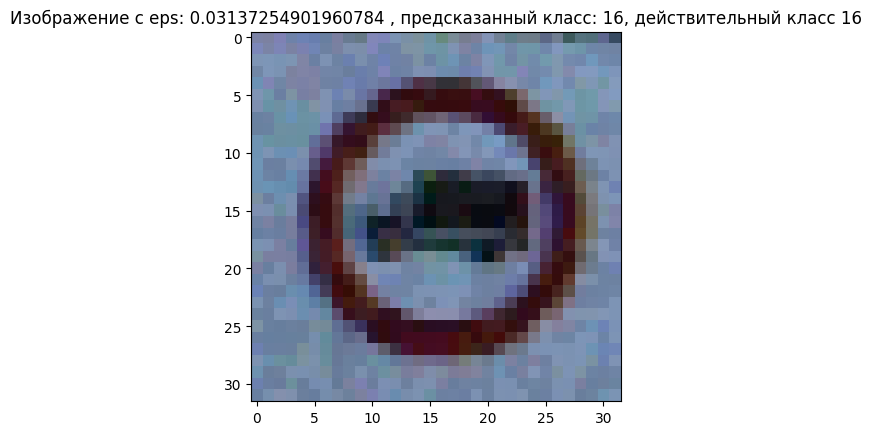
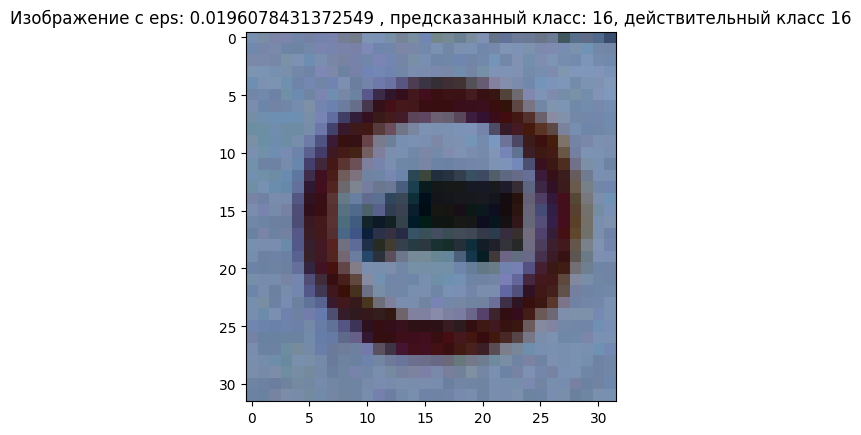
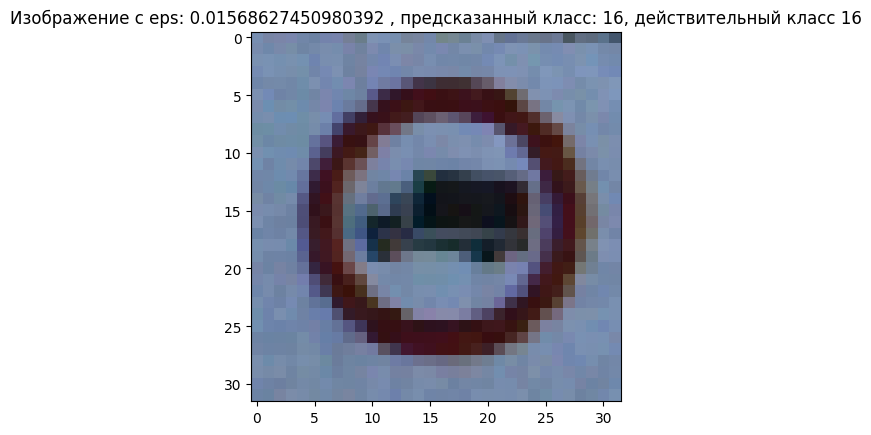
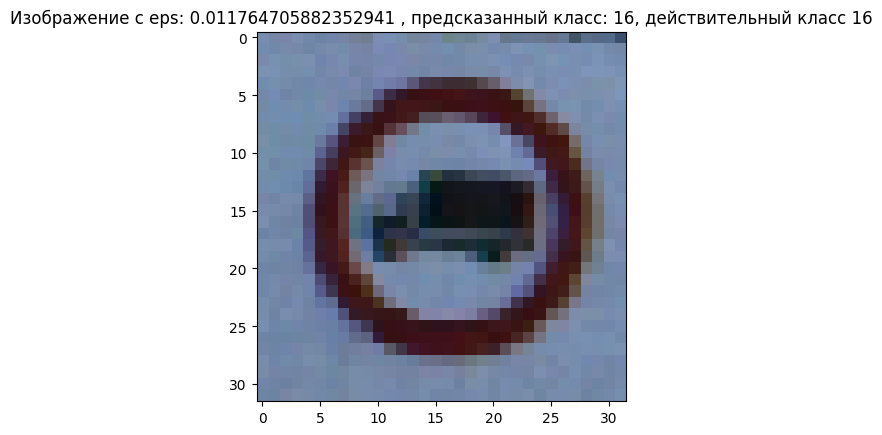
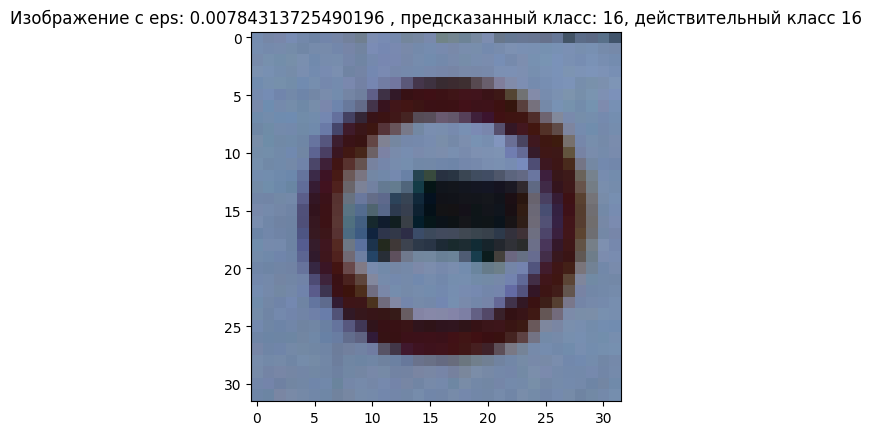
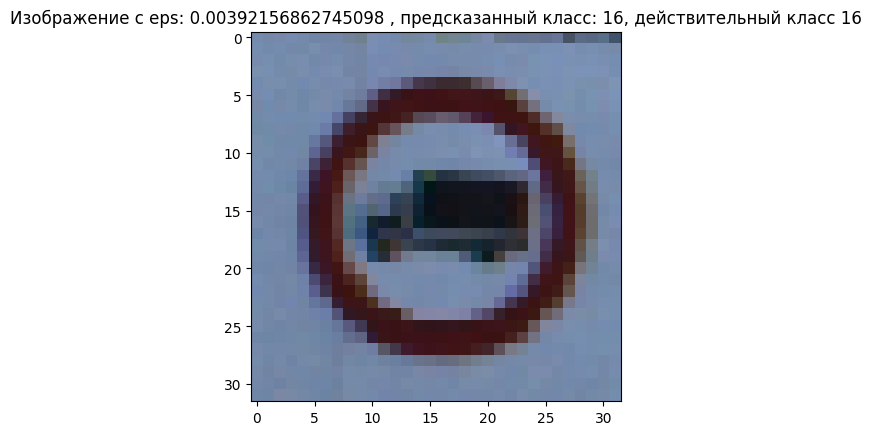
Графики точности для атак на RESNET50



Аналогично проведем PGD и FGSM атаки на VGG16



Пример изображений с различным значением eps для VGG16



Атака PGD на VGG16

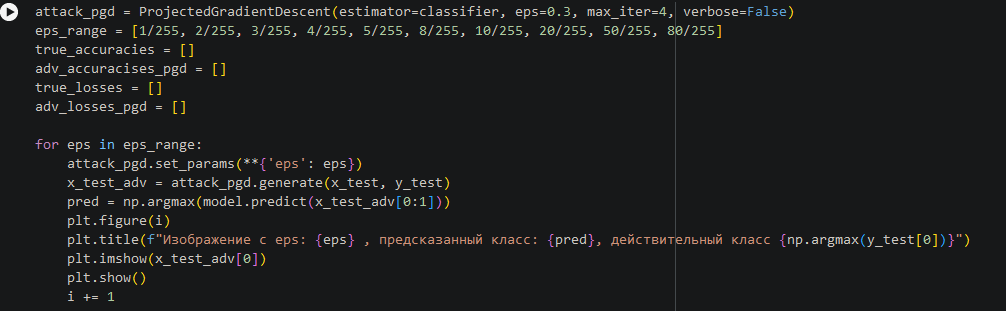
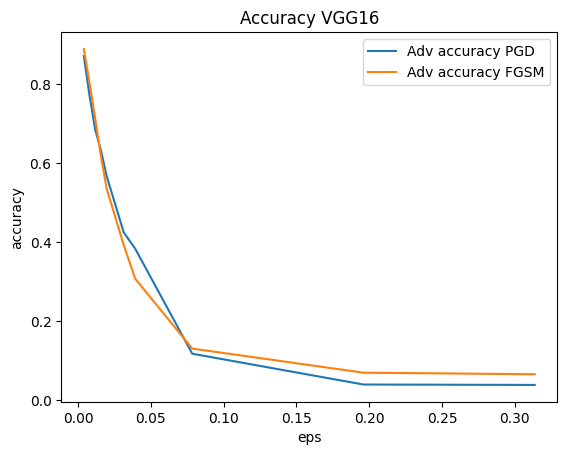
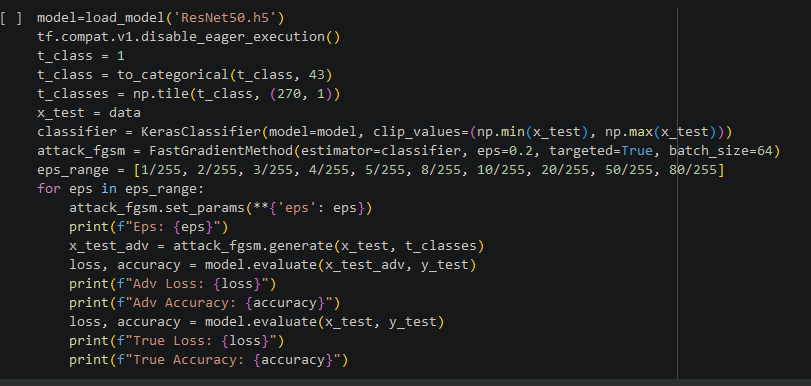


График точности для FGSM и PGD для VGG16



Проведем целевую FGSM атаку

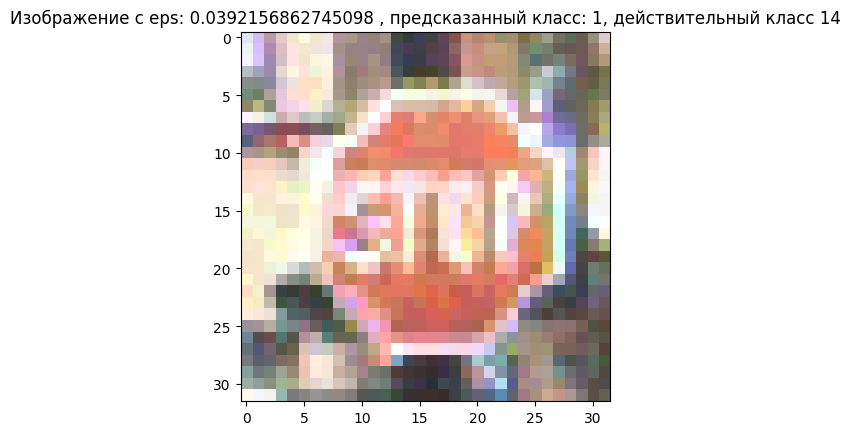


Примеры изображений и результатов классификации

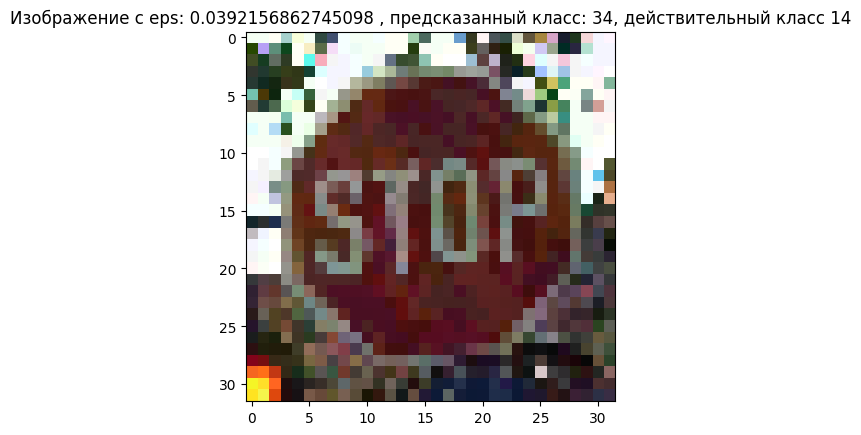








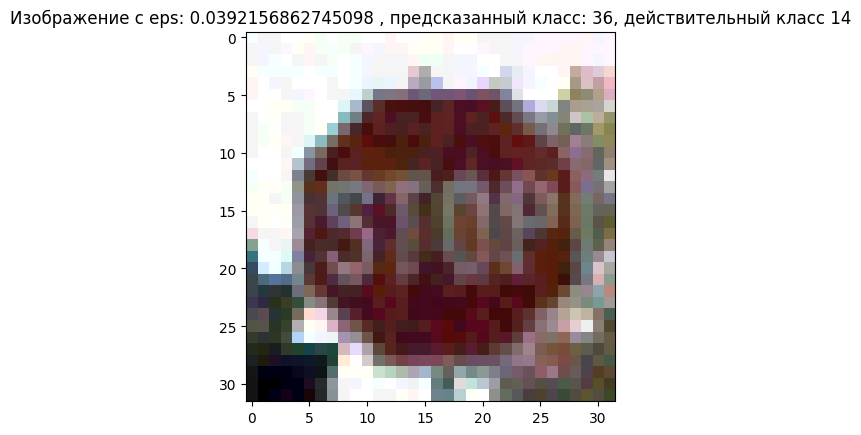




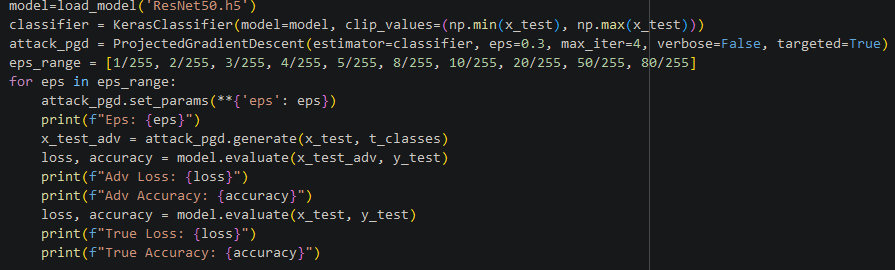






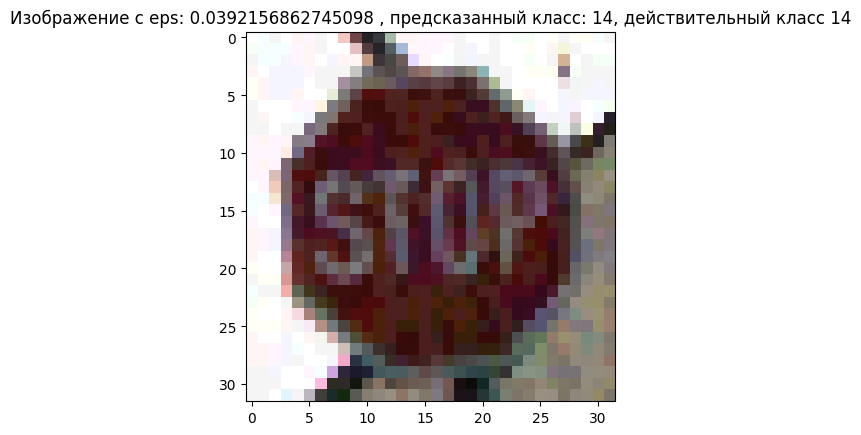


Проведем целевую PGD атаку

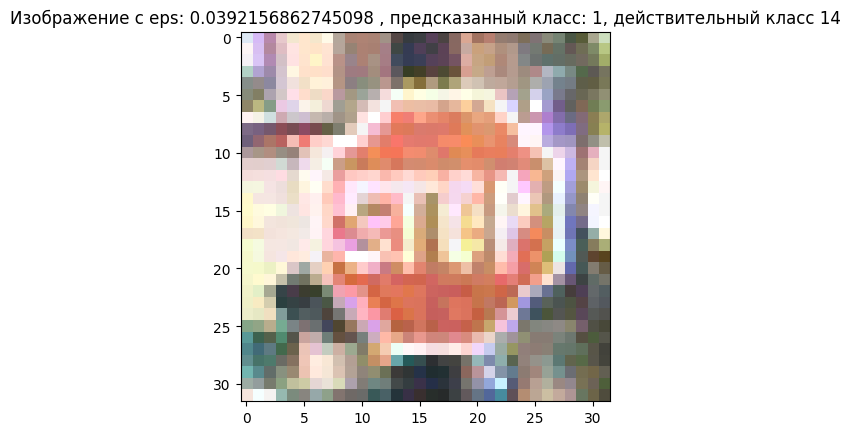


Примеры изображений и результатов классификации

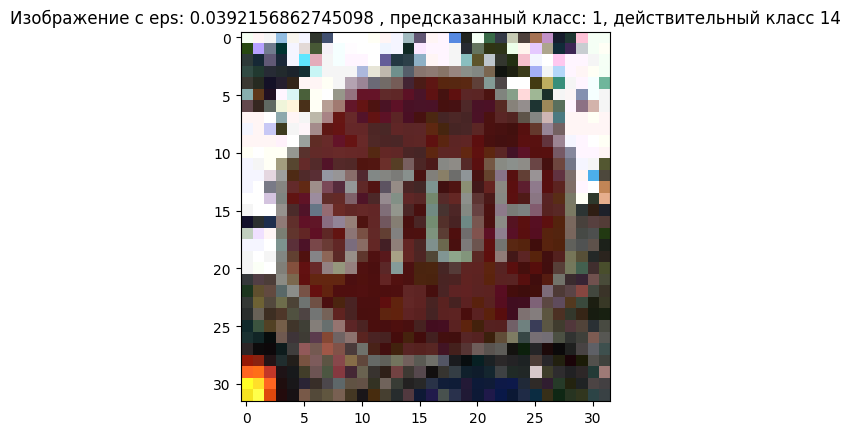




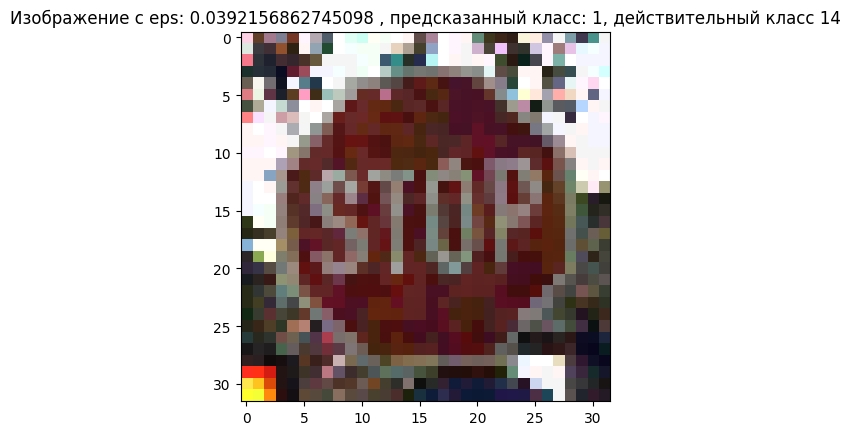














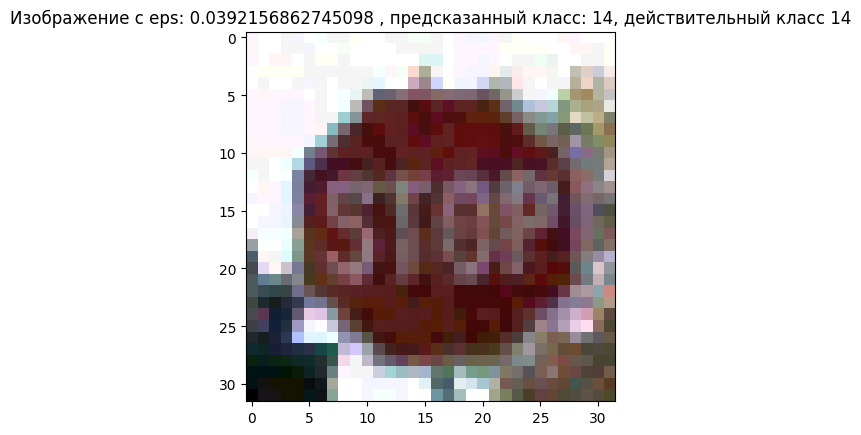
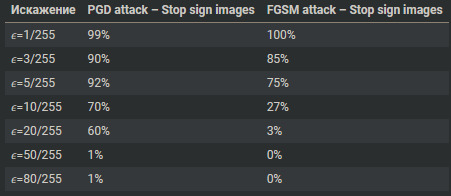


Таблица результатов для целевых FGSM и PGD атак



Метод FGSM не рекомендуется для целевых атак, так как с увеличением значения eps и соответствующего шума возникают ошибки в классификации. Однако, при использовании eps, равного 10/255, модель все же более точно определяет наш целевой класс. При больших значениях eps модель определяет неправильные значения, несоответствующие нашим указаниям (не label 1).

Метод PGD идеально подходит для целевых атак. При больших значениях eps модель почти всегда будет определять класс 14 как класс 1, однако изображение будет слишком зашумленным. Для данного типа атаки оптимальным значением eps будет 50/255. Хотя такие значения сильно зашумляют изображение, классификация класса 1 как класса 14 будет наиболее явно выраженной.

Вывод:

В ходе выполнения 2 лабораторной работы было создано два классификатора, основанных на глубоких нейронных сетях, с использованием датасета GTSRB. Была оценена и представлена точность и графики точности, потерь и валидации для этих классификаторов. Также была применена нецелевая атака уклонения на модели глубокого обучения, основывающаяся на методах белого ящика, и были показаны искаженные изображения. Была составлена таблица с результатами задания 2. Далее была применена целевая атака уклонения на модели глубокого обучения с использованием методов белого ящика, и результаты атаки были представлены в финальной таблице.