

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.

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>

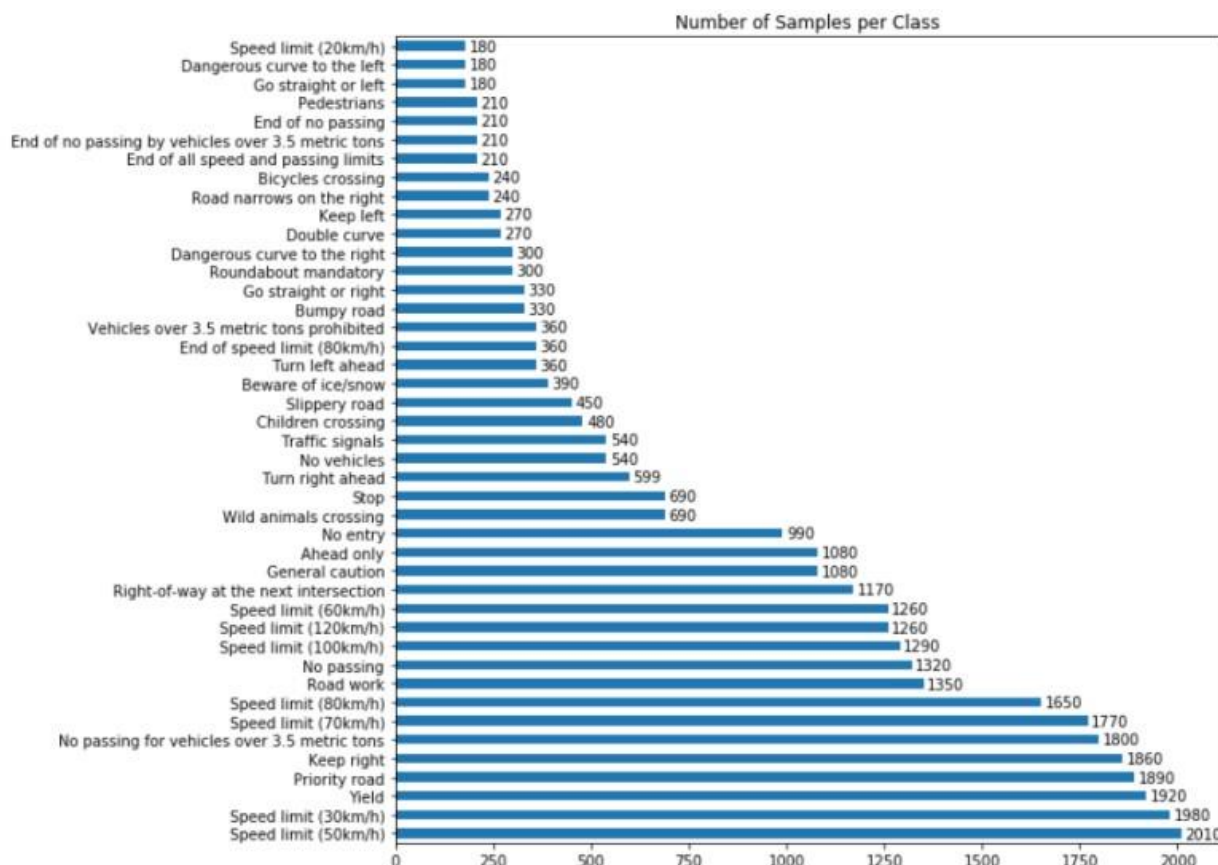


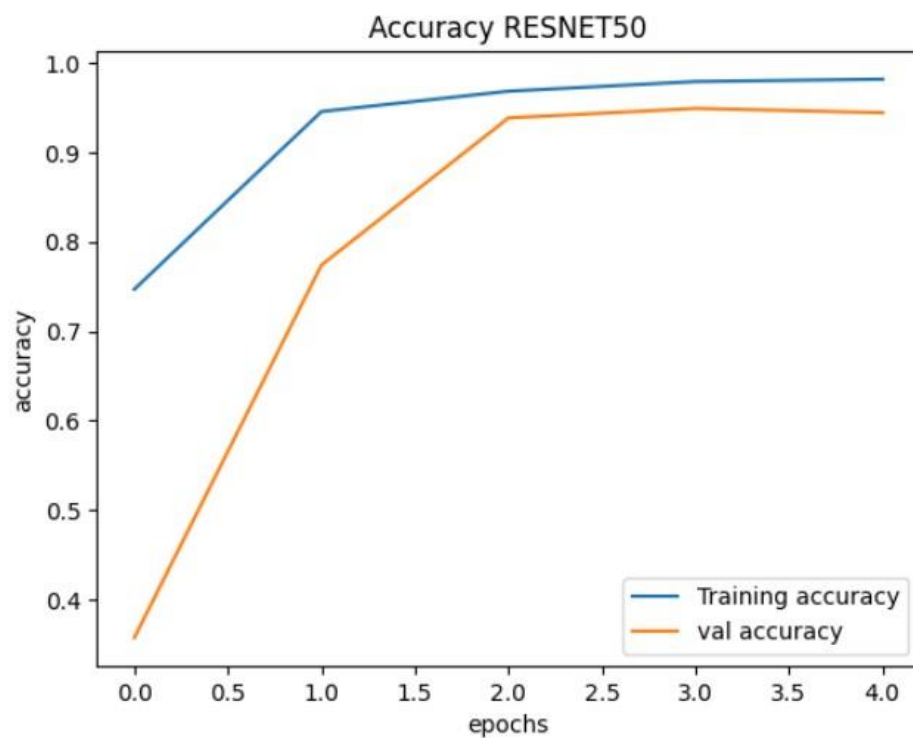
Рис. 1. Распределение изображений в GTSRB

Задание 1

Создаем модель ResNet50, выборки поделены 70/30

```
x_train, x_val, y_train, y_val = train_test_split(data, labels, test_size=0.3, random_state=1)
img_size = (224,224)
model = Sequential()
model.add(ResNet50(include_top = False, pooling = 'avg'))
model.add(Dropout(0.1))
model.add(Dense(256, activation="relu"))
model.add(Dropout(0.1))
model.add(Dense(43, activation = 'softmax'))
model.layers[2].trainable = False
```

Первый график отображает точность обучения и валидации модели RESNET50



Второй график отображает потерю обучения и валидации модели RESNET50

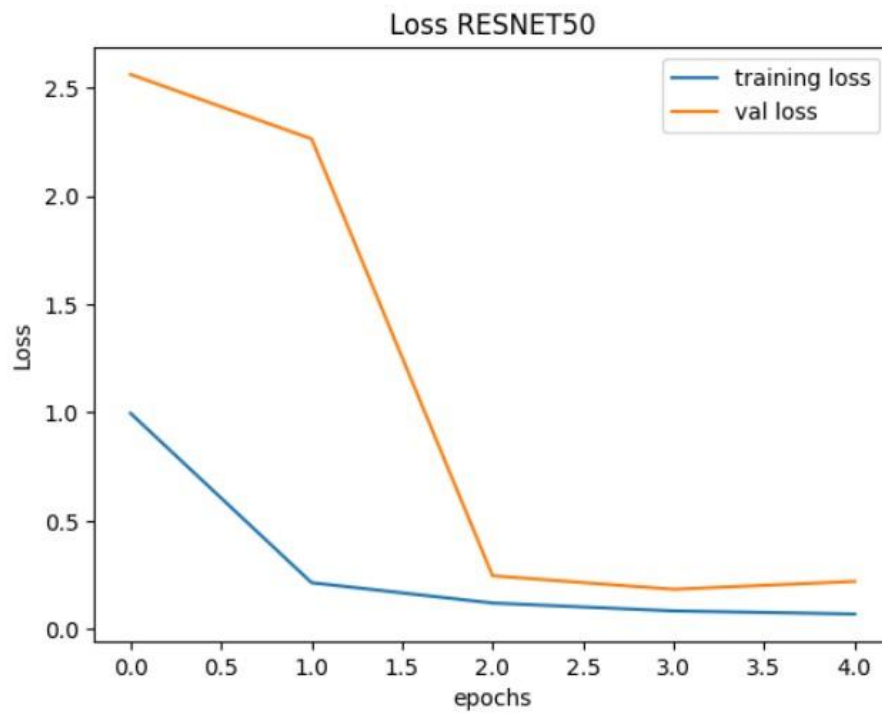
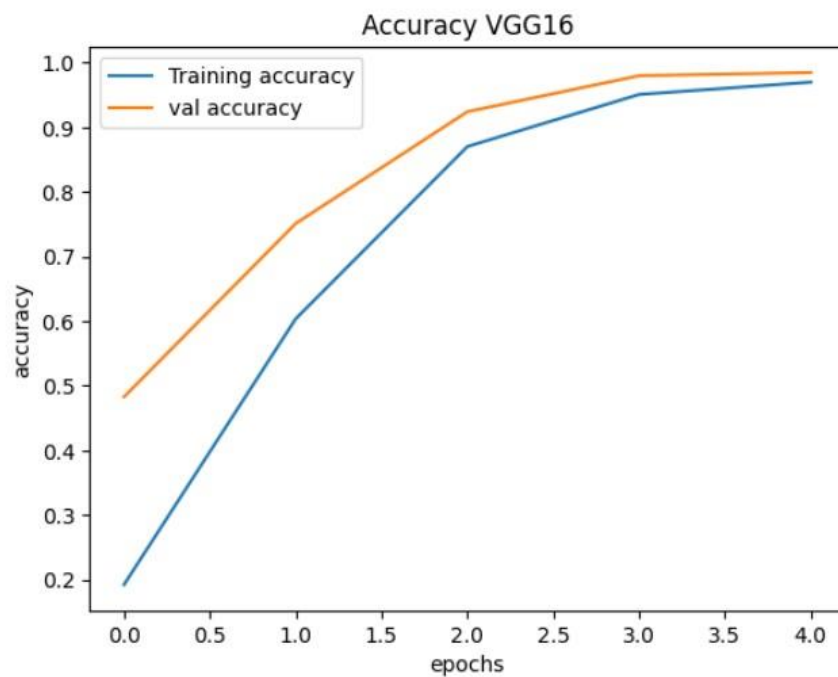


Рис.7. Loss ResNet50

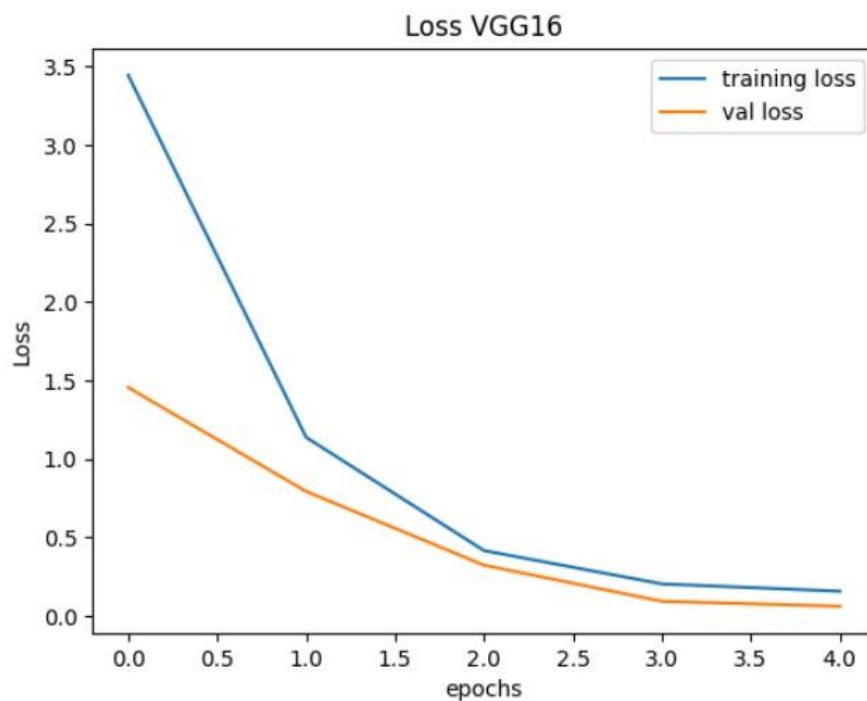
Создаем модель VGG16

```
img_size = (224,224)
model = Sequential()
model.add(VGG16(include_top=False, pooling = 'avg'))
model.add(Dropout(0.1))
model.add(Dense(256, activation="relu"))
model.add(Dropout(0.1))
model.add(Dense(43, activation = 'softmax'))
model.layers[2].trainable = False
```

Третий график отображает точность обучения и валидации модели VGG16



Четвертый график отображает потерю обучения и валидации модели VGG16



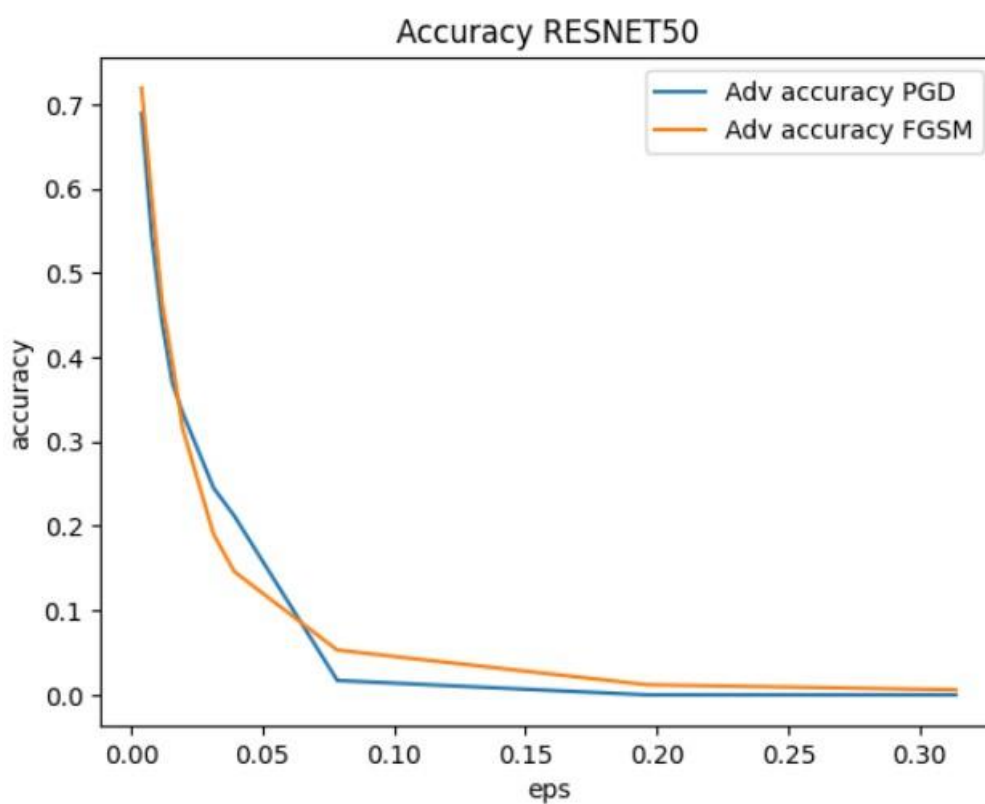
Модель	Обучение	Валидация	Тест
ResNet50	loss: 0.0697 accuracy: 0.9816	loss: 0.2205 accuracy: 0.9442	loss: 0.4797 accuracy: 0.8907
VGG16	loss: 0.1551 accuracy: 0.9698	loss: 0.0592 accuracy: 0.9847	loss: 0.2825 accuracy: 0.9426

Задание 2

Проведем атаки FGSM и PGD на модель RESNET50, используя первые 1,000 изображений из тестового множества. Используем значения параметра искажения:

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

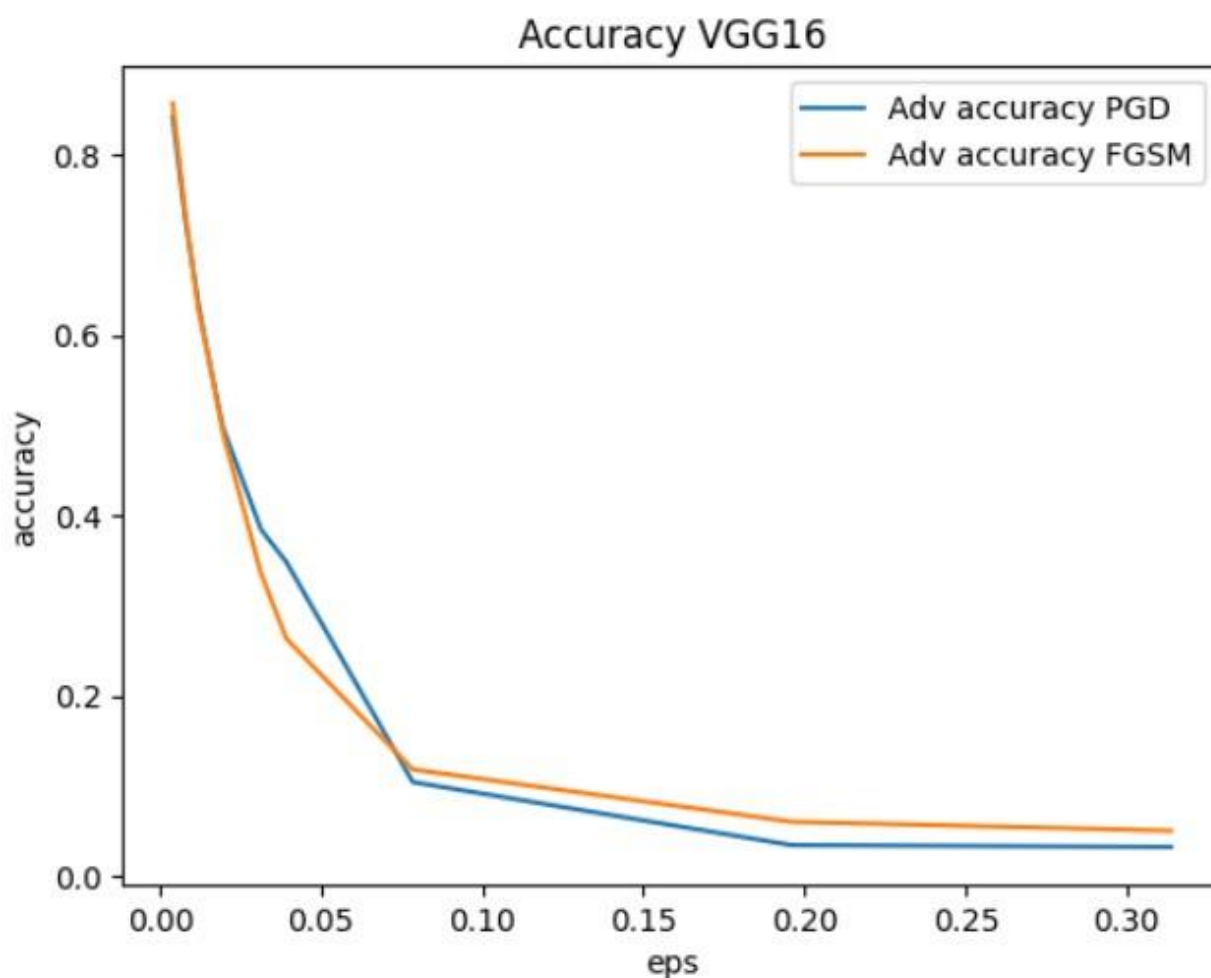
Построим график зависимости точности классификации от параметра искажений эпсилон для RESNET50



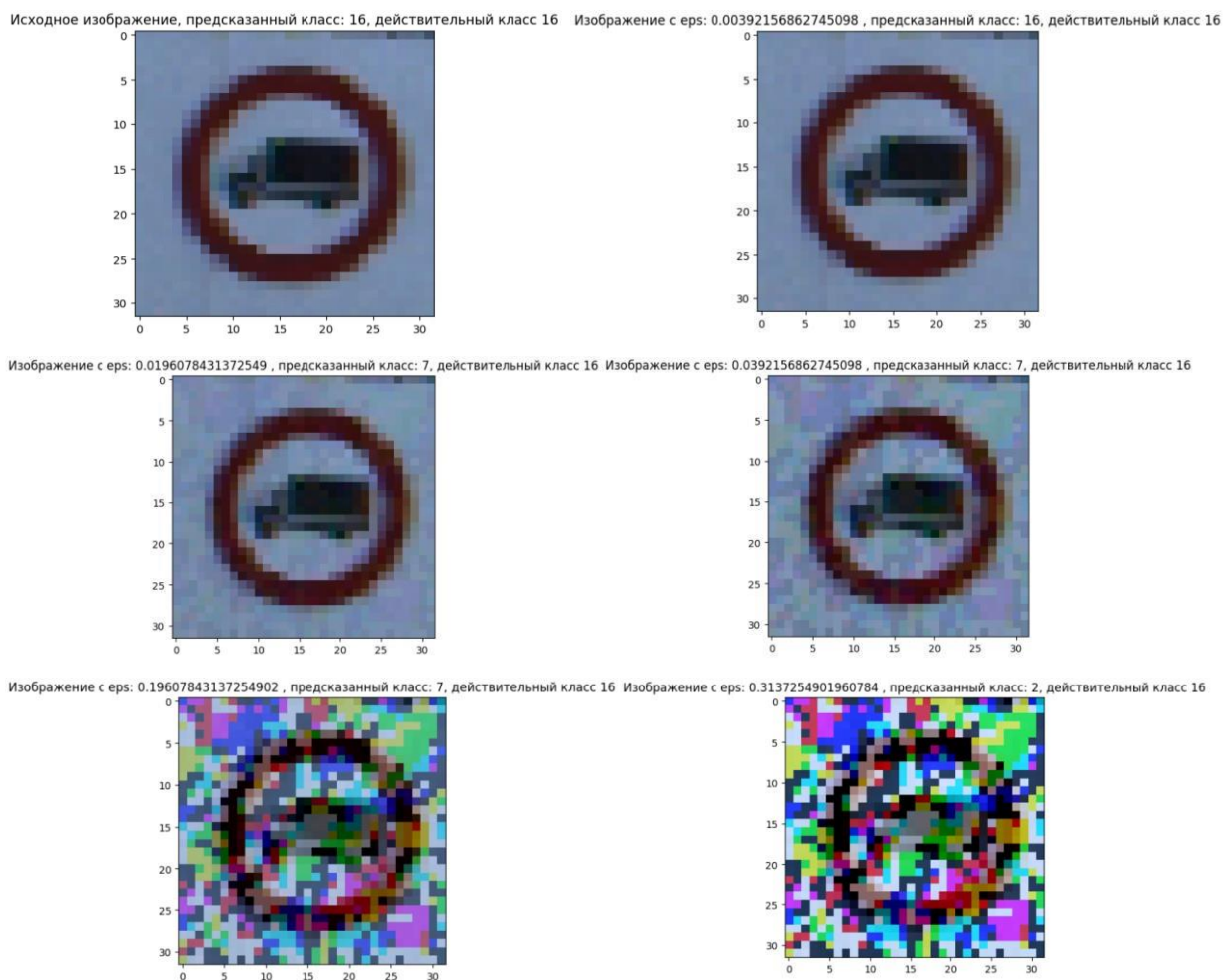
Проведем атаки FGSM и PGD на модель VGG16, используя первые 1,000 изображений из тестового множества. Используем значения параметра искажения:

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

Построим график зависимости точности классификации от параметра искажений эпсилон для VGG16

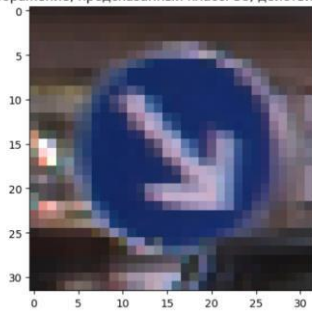


Для атаки FGSM RESNET50, отобразим исходное изображение из датасета и атакующее изображение с указанием величины параметра $\epsilon \in [1/255, 5/255, 10/255, 50/255, 80/255]$, также отобразим предсказанный класс атакующего изображения

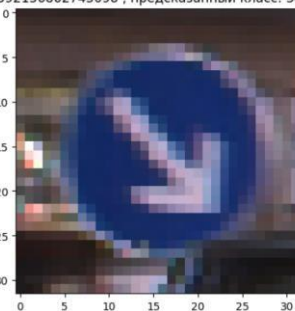


Для атаки FGSM VGG16, отобразим исходное изображение из датасета и атакующее изображение с указанием величины параметра $\epsilon \in [1/255, 5/255, 10/255, 50/255, 80/255]$, также отобразим предсказанный класс атакующего изображения

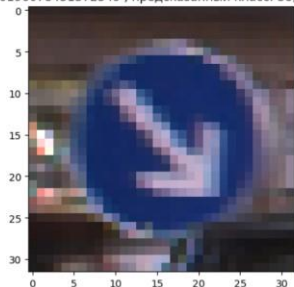
Исходное изображение, предсказанный класс: 38, действительный класс 38



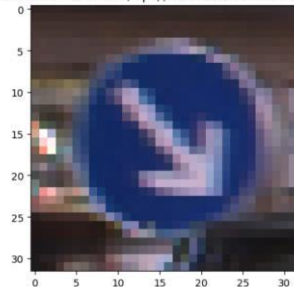
Изображение с eps: 0.00392156862745098 , предсказанный класс: 38, действительный класс 38



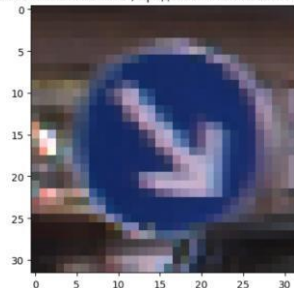
Изображение с eps: 0.0196078431372549 , предсказанный класс: 38, действительный класс 38



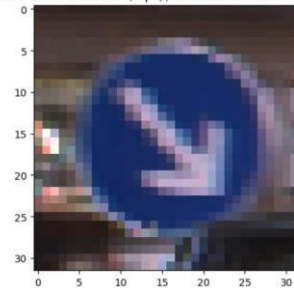
Изображение с eps: 0.0392156862745098 , предсказанный класс: 38, действительный класс 38



Изображение с eps: 0.19607843137254902 , предсказанный класс: 38, действительный класс 38



Изображение с eps: 0.3137254901960784 , предсказанный класс: 38, действительный класс 38



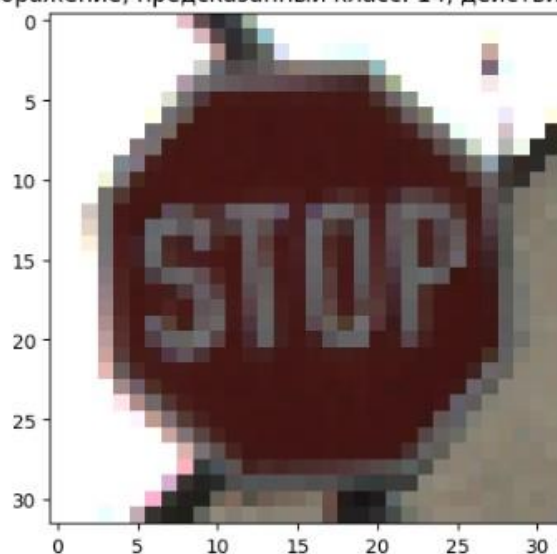
Модель	Исходные изображения	Adversarial images $\epsilon=1/255$	Adversarial images $\epsilon=5/255$	Adversarial images $\epsilon=10/255$
VGG16 - FGSM	89%	79%	44%	21%
VGG16 - PGD	89%	77%	48%	32%
ResNet50 - FGSM	91%	74%	33%	17%
ResNet50 - PGD	91%	71%	30%	23%

Задание 3

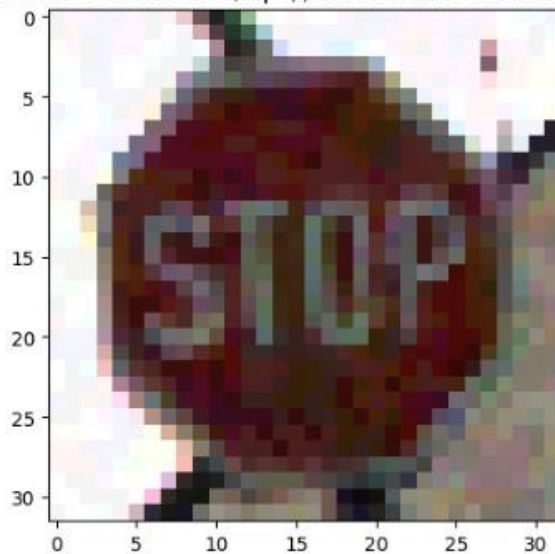
Используя изображения знака «Стоп» (label class 14) из тестового набора данных, применим атаки FGSM и PGD на знак «Стоп» с целью классификации его как знака «Ограничение скорости 30» (target label class = 1), изменяя значения искажений $\epsilon = [1/255, 3/255, 5/255, 10/255, 20/255, 50/255, 80/255]$.

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

Исходное изображение, предсказанный класс: 14, действительный класс 14



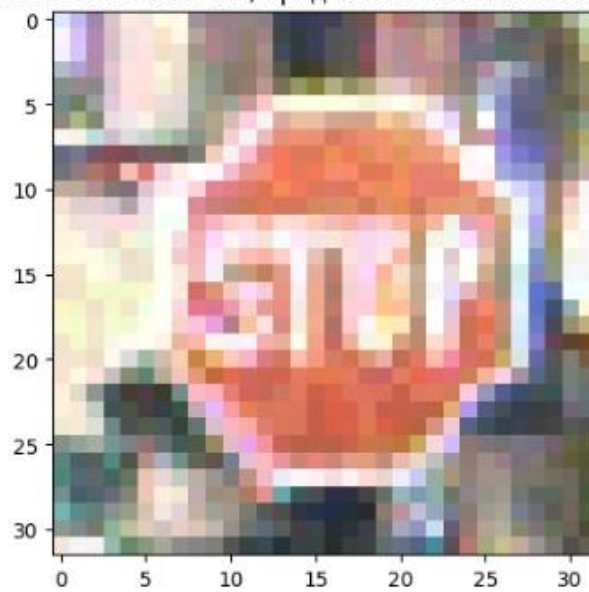
Изображение с eps: 0.0392156862745098 , предсказанный класс: 24, действительный класс 14



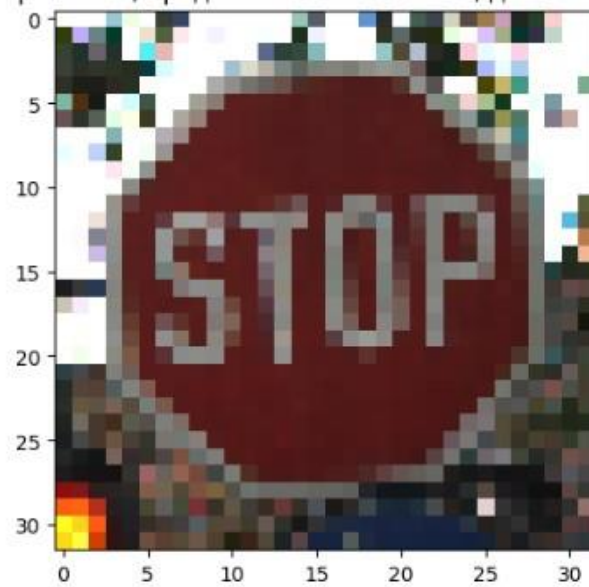
Исходное изображение, предсказанный класс: 11, действительный класс 14



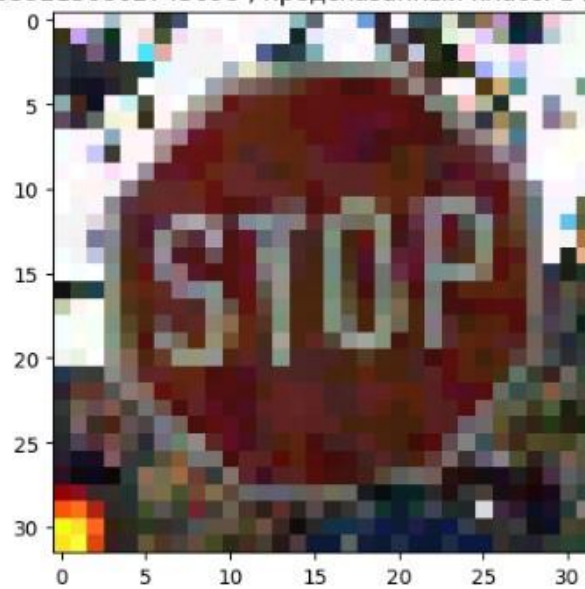
Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



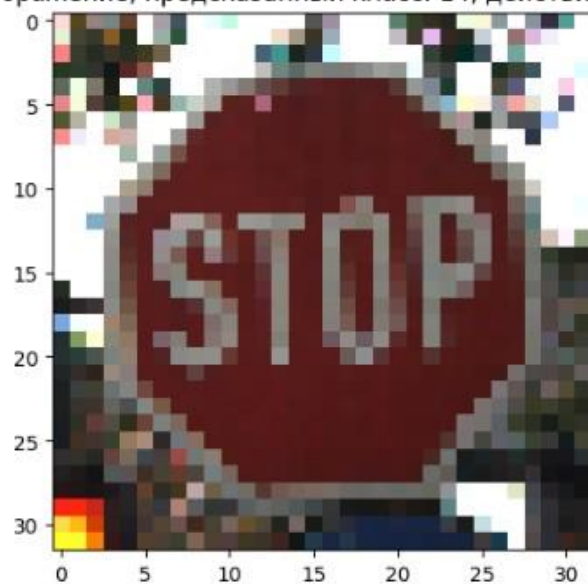
Исходное изображение, предсказанный класс: 14, действительный класс 14



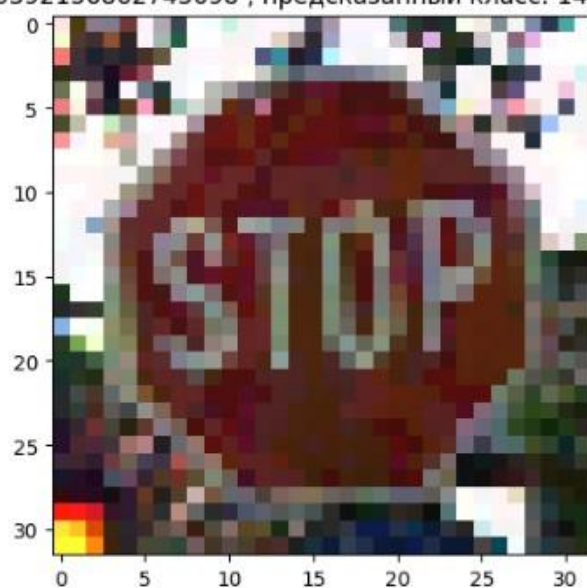
Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



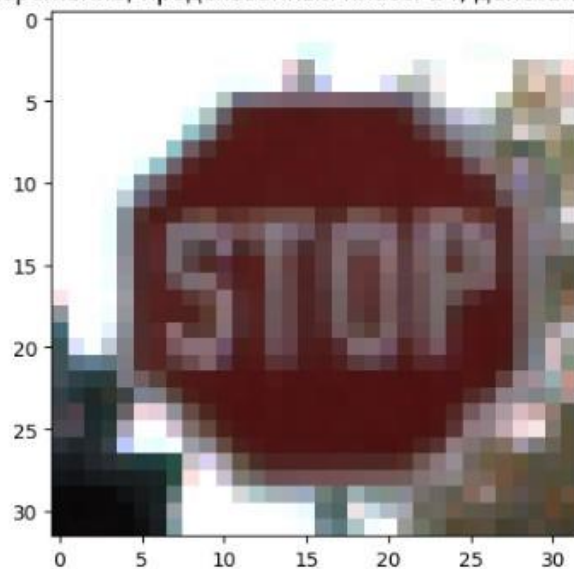
Исходное изображение, предсказанный класс: 14, действительный класс 14



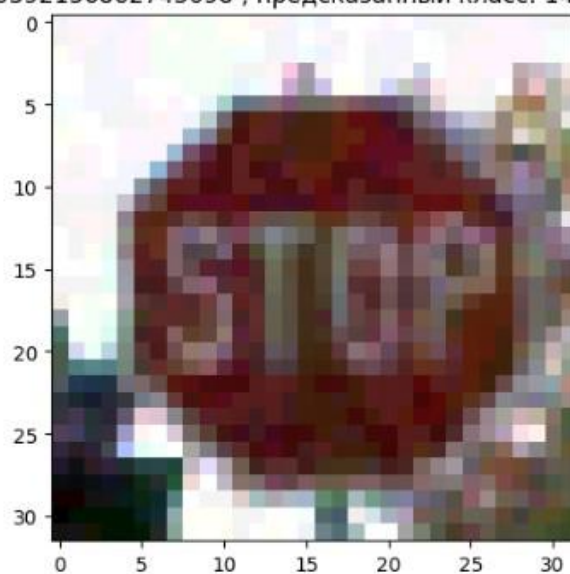
Изображение с ерс: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



Исходное изображение, предсказанный класс: 14, действительный класс 14

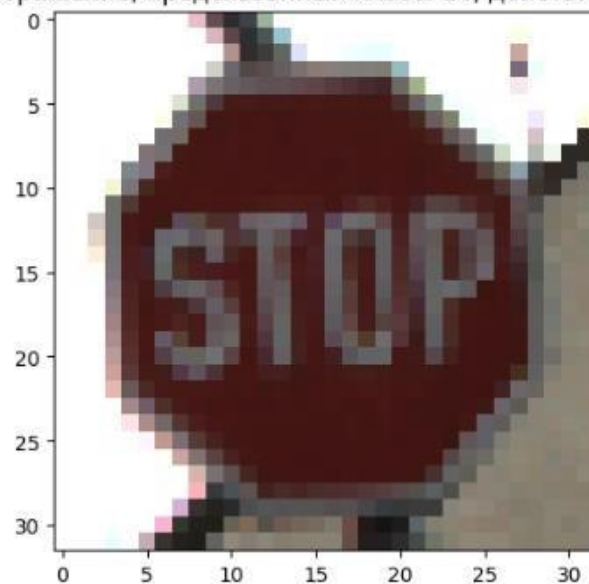


Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14

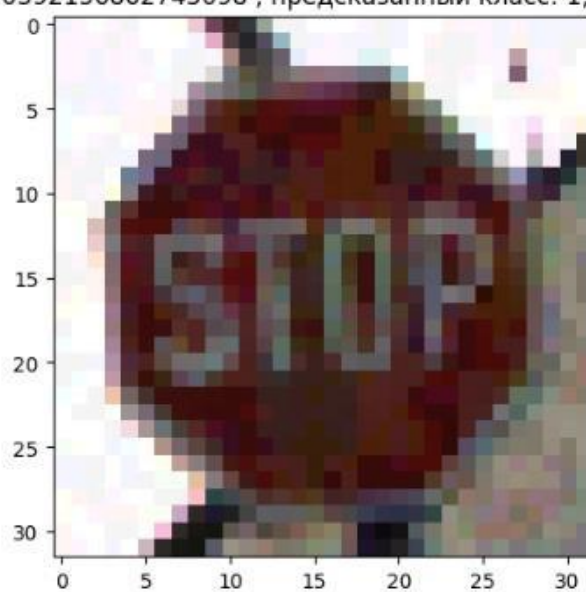


Выведем 5 пар примеров исходных изображений знака «Стоп» и соответствующих атакующих примеров для атаки PG

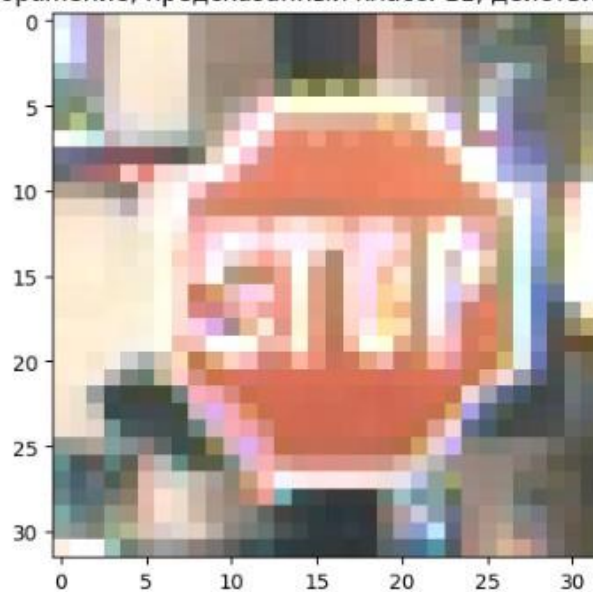
Исходное изображение, предсказанный класс: 14, действительный класс 14



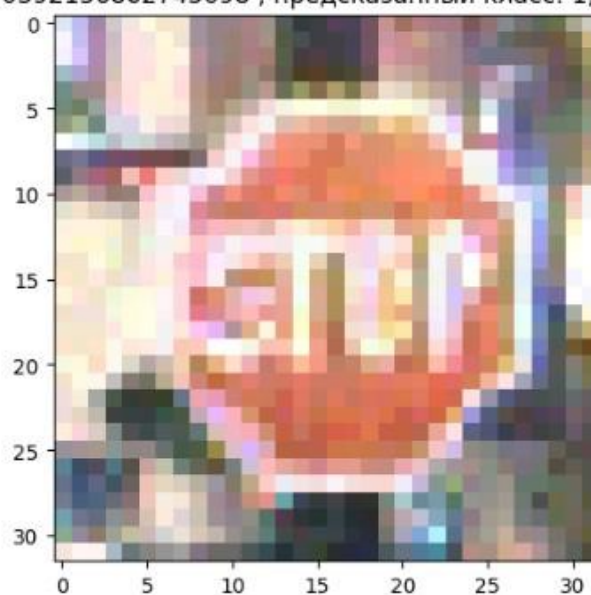
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



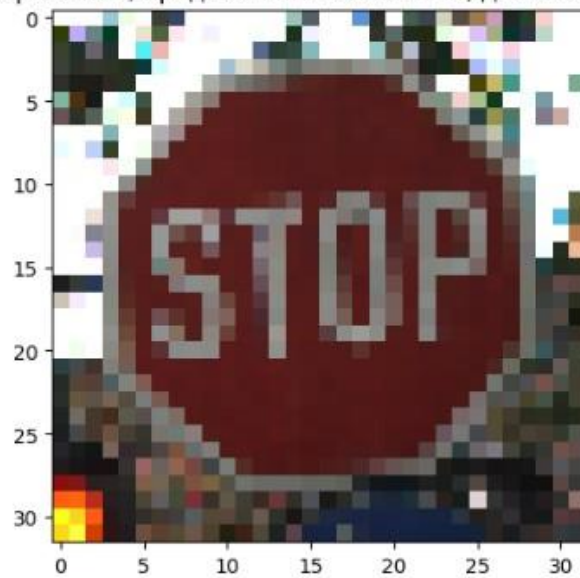
Исходное изображение, предсказанный класс: 11, действительный класс 14



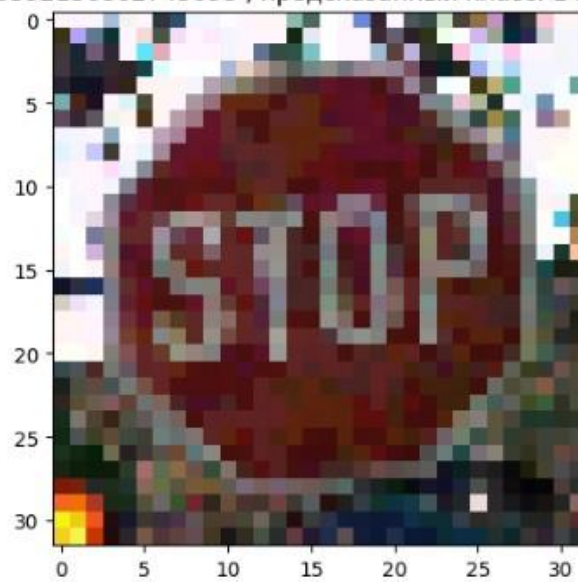
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



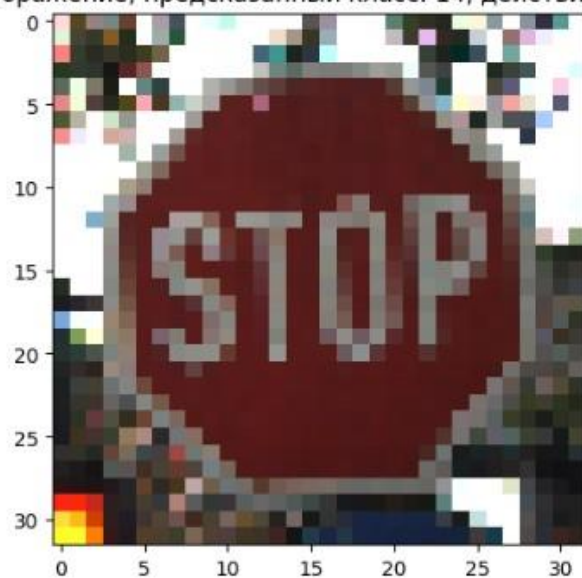
Исходное изображение, предсказанный класс: 14, действительный класс 14



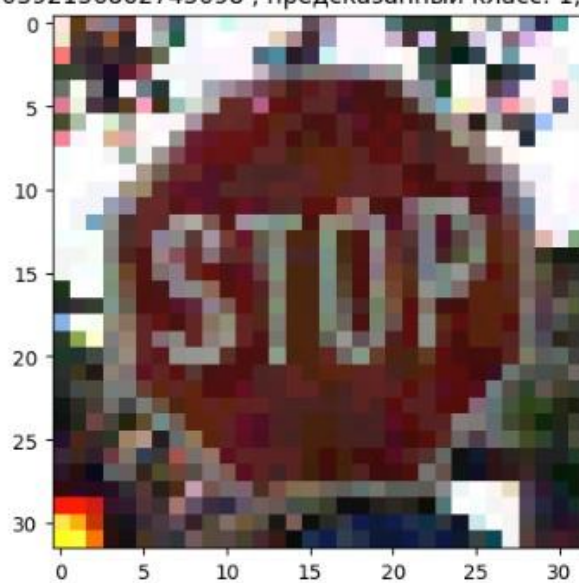
Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



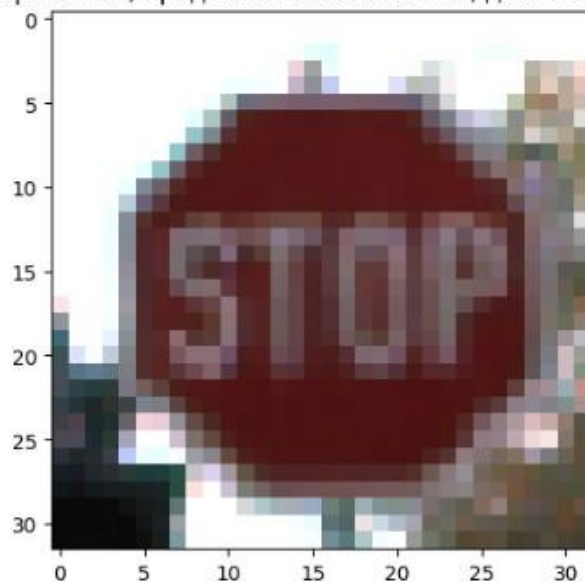
Исходное изображение, предсказанный класс: 14, действительный класс 14



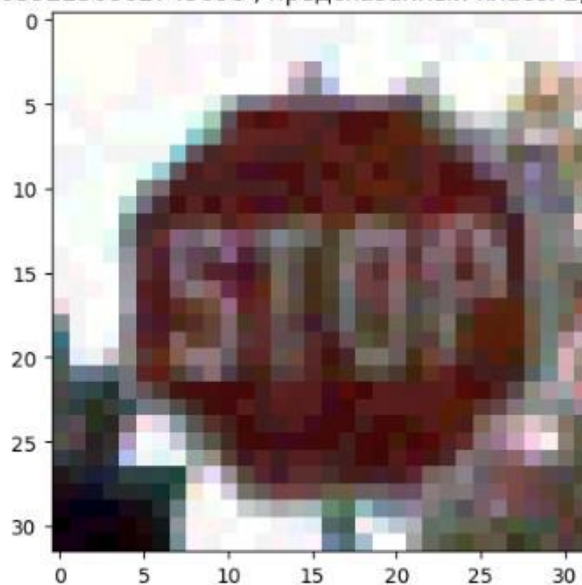
Изображение с ерс: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



Исходное изображение, предсказанный класс: 14, действительный класс 14



Изображение с eps: 0.0392156862745098 , предсказанный класс: 2, действительный класс 14



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.