МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

ИКБ направление «Кибербразведка и противодействие угрозам с применением технологий искусственного интеллекта» 10.04.01

Кафедра КБ-4 «Интеллектуальные системы информационной безопасности»

Лабораторная работа №2

по дисциплине

«Анализ защищенности систем искусственного интеллекта»

Группа:
ББМО-01-22
Выполнил:
Некрасов Е.А.

Проверил:
Спирин А.А.

Москва 2023

Задание 1

Для этой работы используем набор данных GTSRB (German Traffic Sign Recognition Benchmark). Набор данных состоит примерно из 51 000 изображений дорожных знаков. Загрузим набор данных по ссылке: https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

Обучить 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB. Ресурсы колаба не безграничны, поэтому используем только часть набора данных. Использовали следующие модели нейронных сетей: ResNet50 и VGG16. Будем использовать необходимые фреймворки.

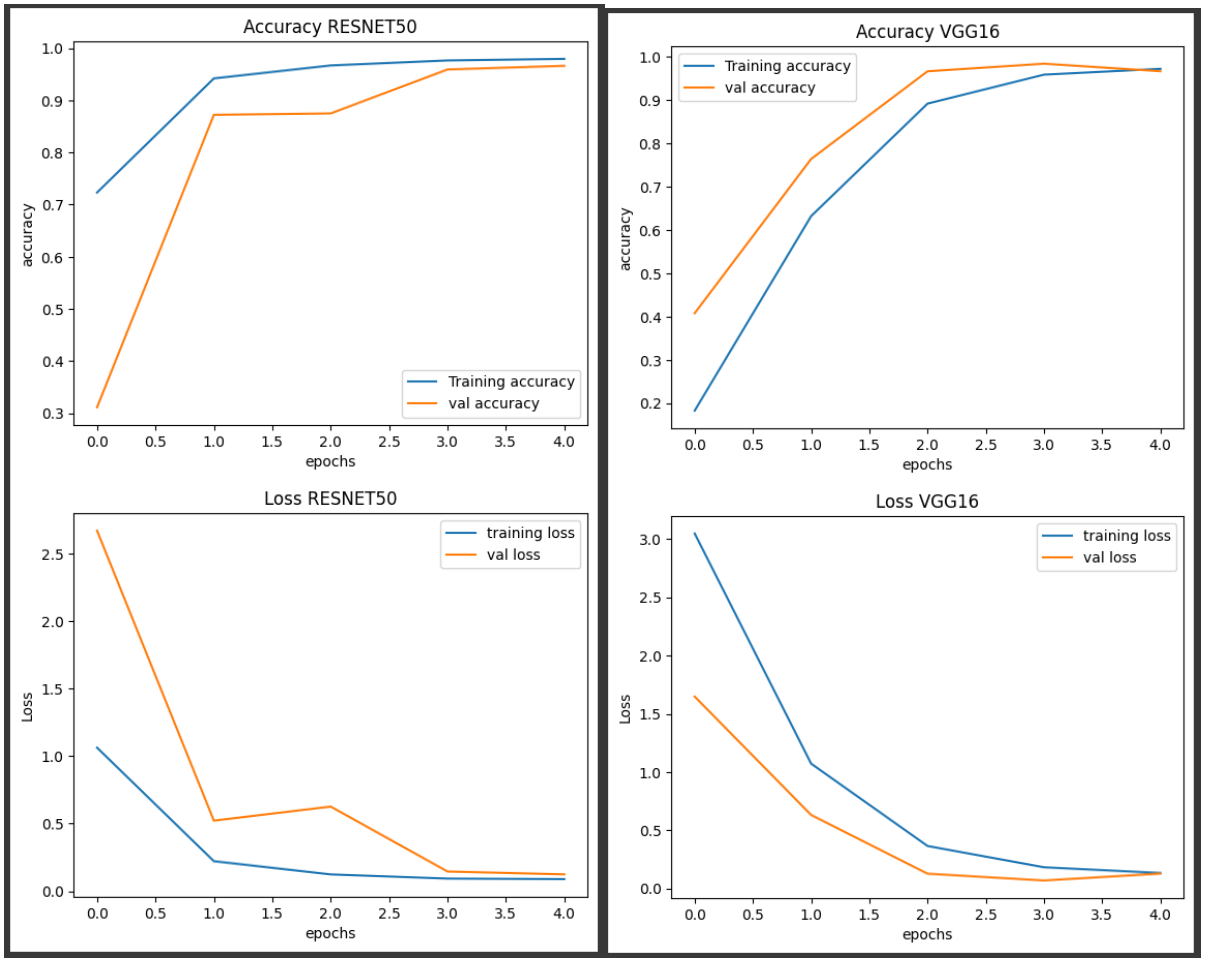
Создадим модель ResNet50:

```
[11] 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

     Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/resnet/resnet50_weights_tf_dim_ordering_tf_kernels_notop.h5
     94765736/94765736 [==============================] - 3s 0us/step
```

Создадим модель VGG16:

```
[17] del model
     del history
     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

     Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels_notop.h5
     58889256/58889256 [==============================] - 2s 0us/step
```

По завершении обучения были сформированы следующие графики точности для моделей ResNet50, VGG16:

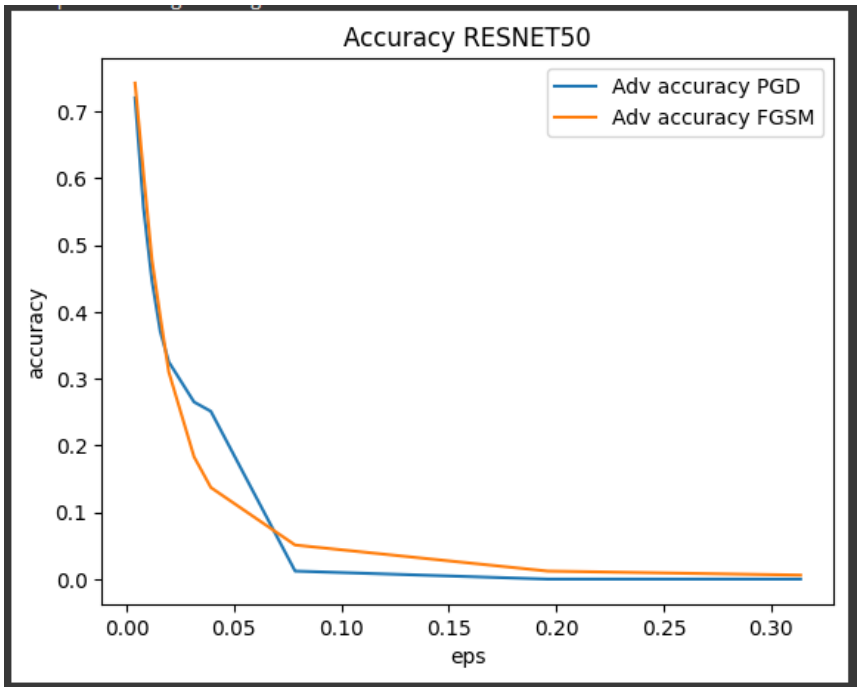


В результате была получена следующая результирующая таблица:

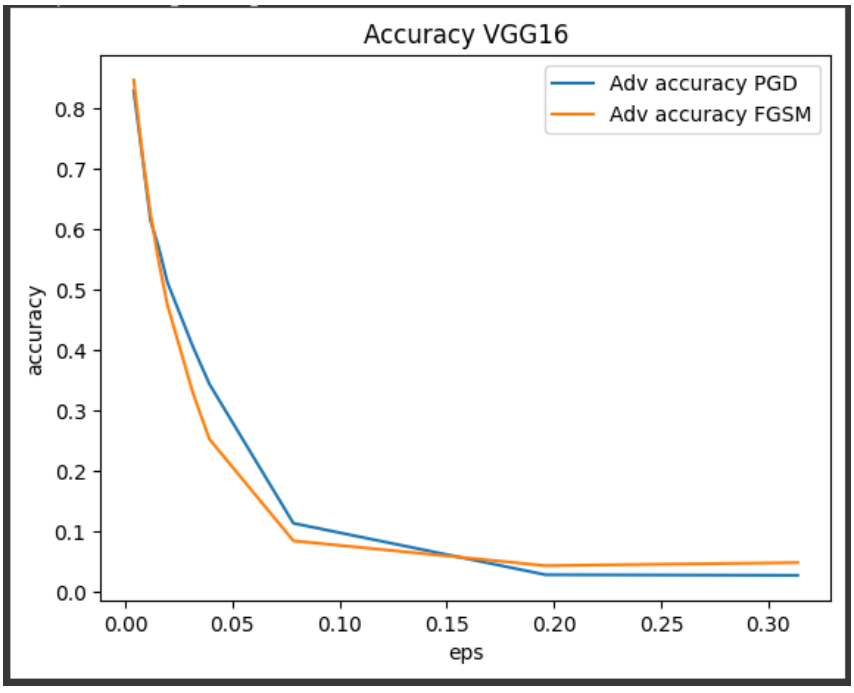
Модель	Обучение	Валидация	Тест
VGG16	Loss:0.1325 accuracy:0.9722	Loss:0.1273 accuracy:0.9667	Loss:0.2896 accuracy:0.9363
ResNet50	Loss:0.089 accuracy:0.9793	Loss:0.1247 accuracy:0.9660	Loss:0.3418 accuracy:0.9209

Задание 2

ResNet50: График зависимости точности классификации от параметра искажения.



VGG16: График зависимости точности классификации от параметра искажения.

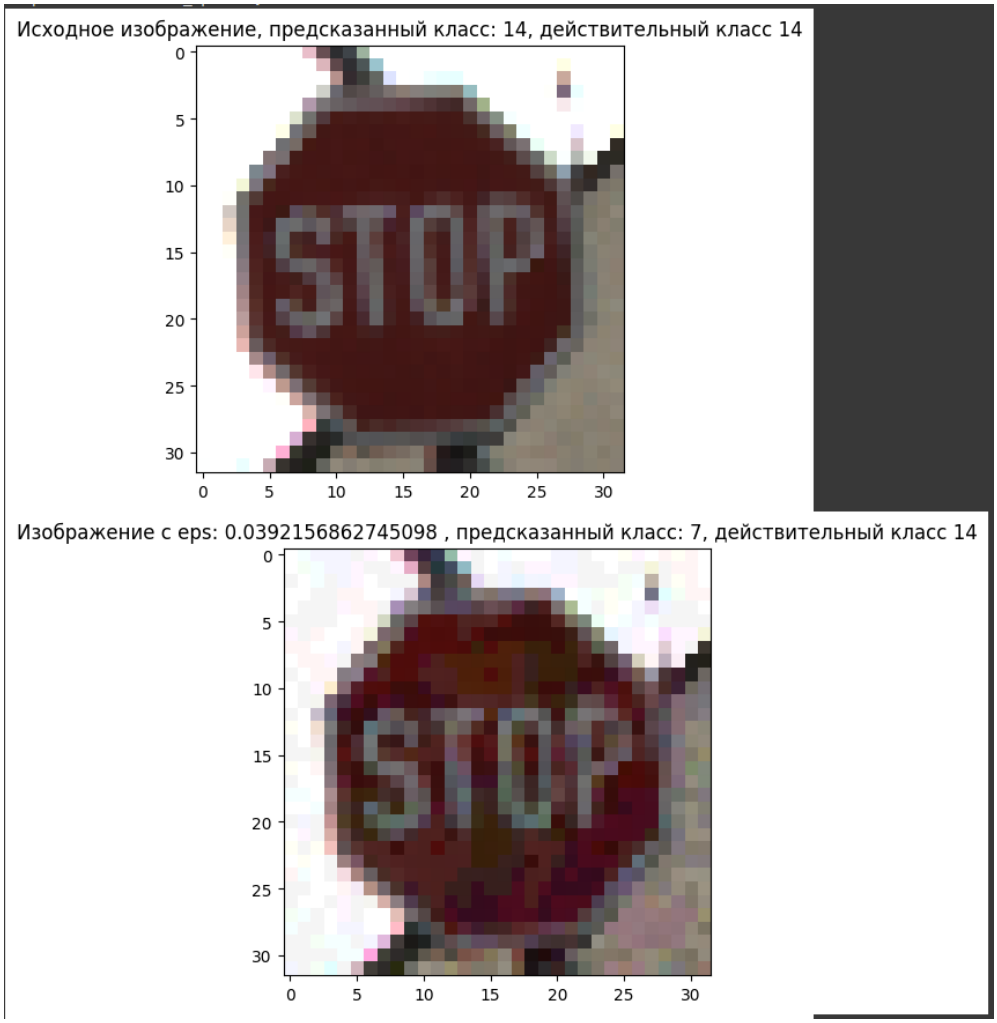


В результате была получена следующая результирующая таблица:

Модель	Исходные изображения	Adversarial images ϵ=1/255	Adversarial images ϵ=5/255	Adversarial images ϵ=10/255
VGG16-FGSM	94%	84%	47%	25%
VGG16-PGD	92%	82%	51%	34%
ResNet50-FGSM	91%	74%	31%	13%
ResNet50-PGD	91%	72%	32%	25%

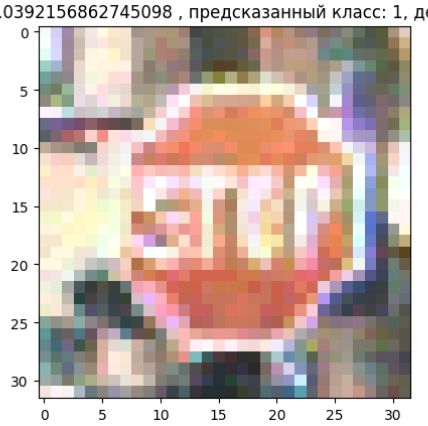
Задание 3

FGSM:Пример исходных изображений знака «Стоп» исоответствующих атакующих примеров.



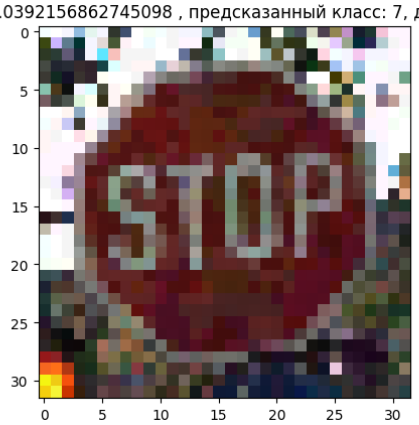
Исходное изображение, предсказанный класс: 14, действительный класс 14

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



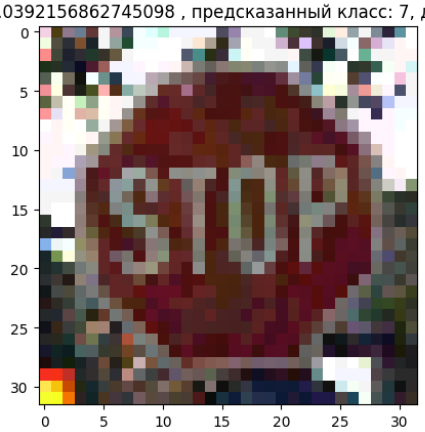
Исходное изображение, предсказанный класс: 14, действительный класс 14

Изображение с eps: 0.0392156862745098 , предсказанный класс: 7, действительный класс 14



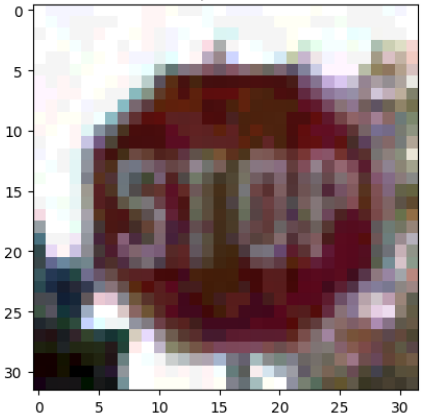
Исходное изображение, предсказанный класс: 14, действительный класс 14

Изображение с eps: 0.0392156862745098 , предсказанный класс: 7, действительный класс 14

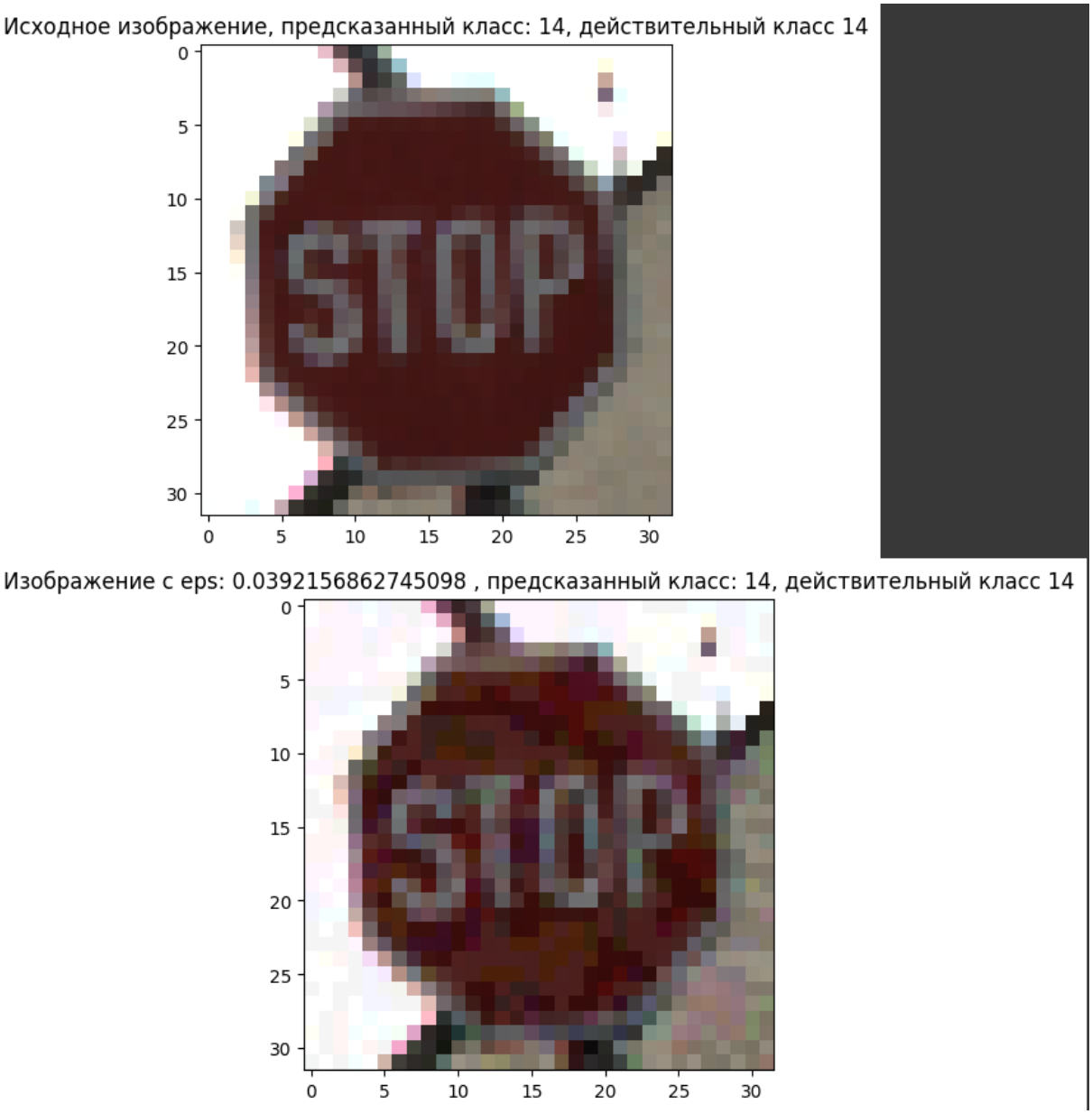


Исходное изображение, предсказанный класс: 14, действительный класс 14

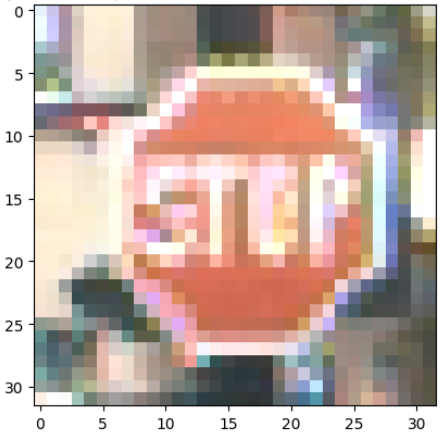
Изображение с eps: 0.0392156862745098 , предсказанный класс: 8, действительный класс 14



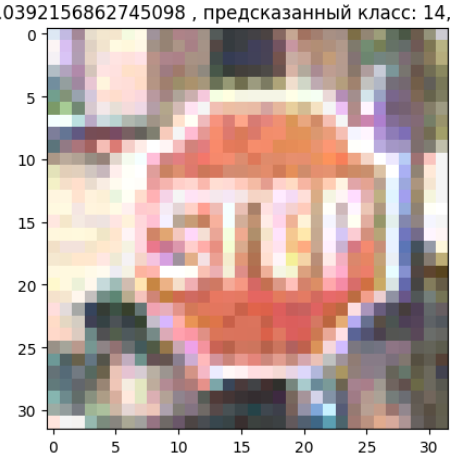
PGD: Пример исходных изображений знака «Стоп» и соответствующих атакующих примеров.



Исходное изображение, предсказанный класс: 14, действительный класс 14

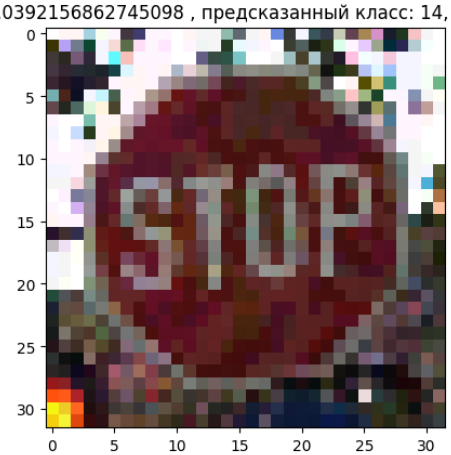


Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



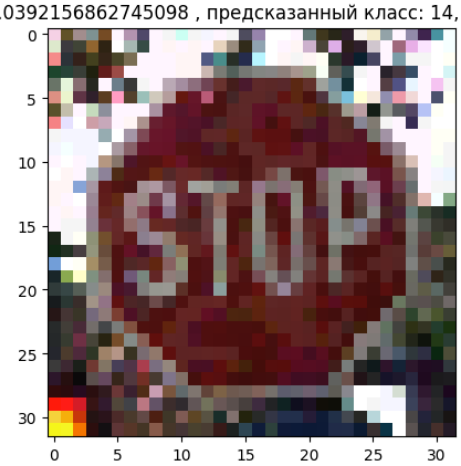
Исходное изображение, предсказанный класс: 14, действительный класс 14

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

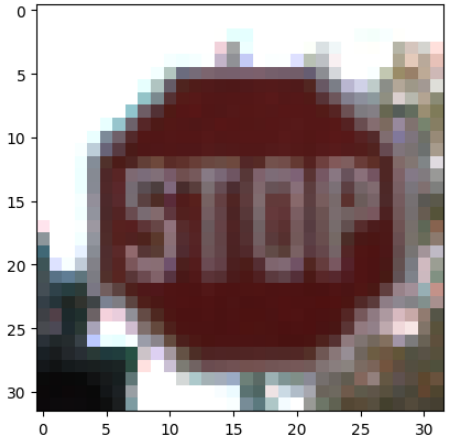


Исходное изображение, предсказанный класс: 14, действительный класс 14

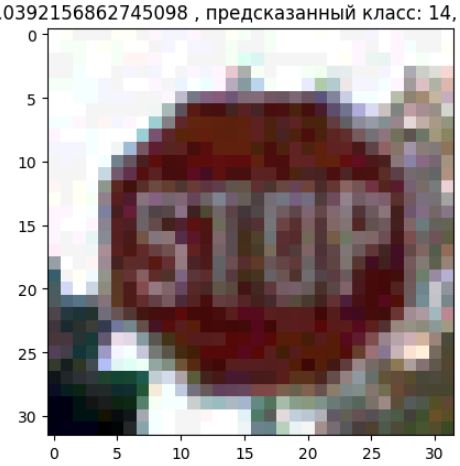
Изображение с eps: 0.0392156862745098 , предсказанный класс: 14, действительный класс 14



Исходное изображение, предсказанный класс: 14, действительный класс 14



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



Результирующая таблица по заданию:

Искажение	PGD attack – Stop sign images	FGSM attack – Stop sign images
ϵ=1/255	96%	91%
ϵ=3/255	92%	68%
ϵ=5/255	76%	47%
ϵ=10/255	72%	1%
ϵ=20/255	26%	0%
ϵ=50/255	0%	0%
ϵ=80/255	0%	0%

Вывод: Метод FGSM плохо подходит для целевых атак. При растущем искажении классификация начинает давать сбой. Для целевой атаки подходит PGD. При больших искажениях, модель почти всегда будет определять заданный нами класс, но изображение станет слишком навязчиво искажено.

В ходе выполнения лабораторной работы были выполнены предоставленные задания:

– Подготовить 2 классификатора на основе глубоких нейронных сетей на датасете GTSRB;

– Применить нецелевую атаку уклонения на основе белого ящика против моделей глубокого обучения;

– Применить целевую атаку уклонения на основе белого ящика против моделей глубокого обучения.