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:
Феденёв А.В.

Проверил:
Спирин А.А.

Москва 2023

Задание 1

В начале работы установили необходимые библиотеки и подключились к google drive.

Создадим модель ResNet50:

```
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
```

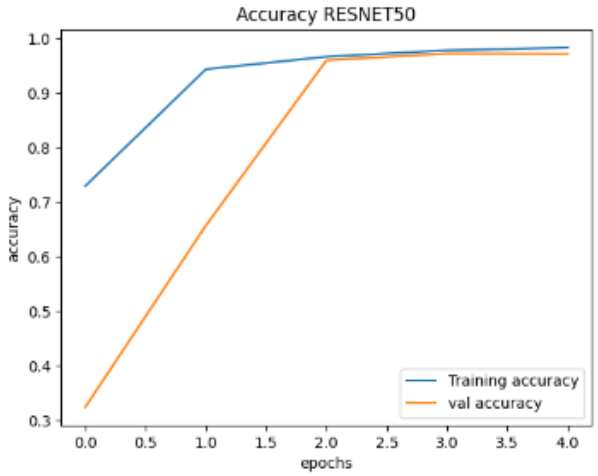
```
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/resnet/resnet50_weights_tf_dim_ordering_tf_kernels_notop.h5
94765736/94765736 [==============================] - 0s 0us/step
```

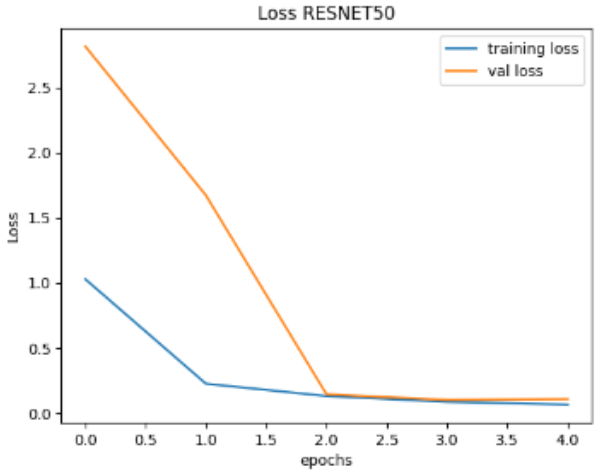
Создадим модель VGG16:

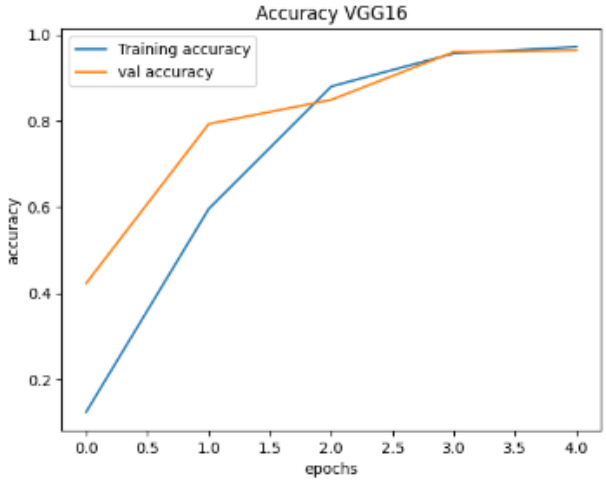
```
img_size = (224,224)
model = Sequential()
model.add(VGG16(include_top=False, pooling = 'avg'))
model.add(Dropout(0.1))
model.add(Dense(256, activation="relu"))
model.add(Dropout(0.1))
model.add(Dense(43, activation = 'softmax'))
model.layers[2].trainable = False
```

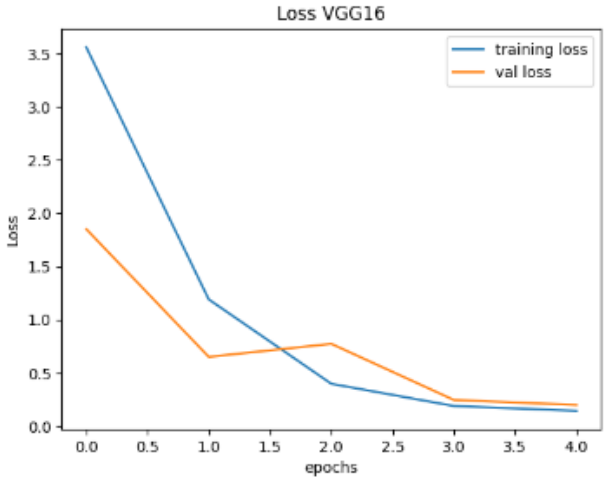
```
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels_notop.h5
58889256/58889256 [==============================] - 0s 0us/step
```

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


Accuracy RESNET50
accuracy
epochs
Training accuracy
val accuracy


Loss RESNET50
Loss
epochs
training loss
val loss





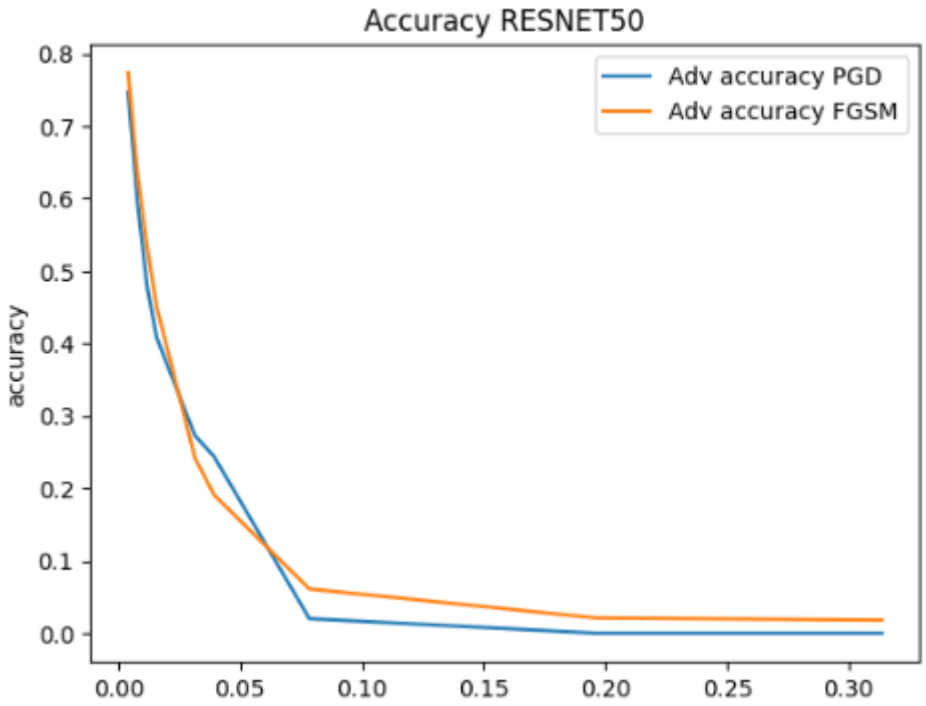
В результате была получена следующая результирующая таблица:

Модель	Обучение	Валидация	Тест
VGG16	Loss:0.1443 accuracy:0.9713	Loss:0.1997 accuracy:0.9631	Loss:0.4049 accuracy:0.9235
ResNet50	Loss:0.065 accuracy:0.9835	Loss:0.1074 accuracy:0.9719	Loss:0.4224 accuracy:0.9172

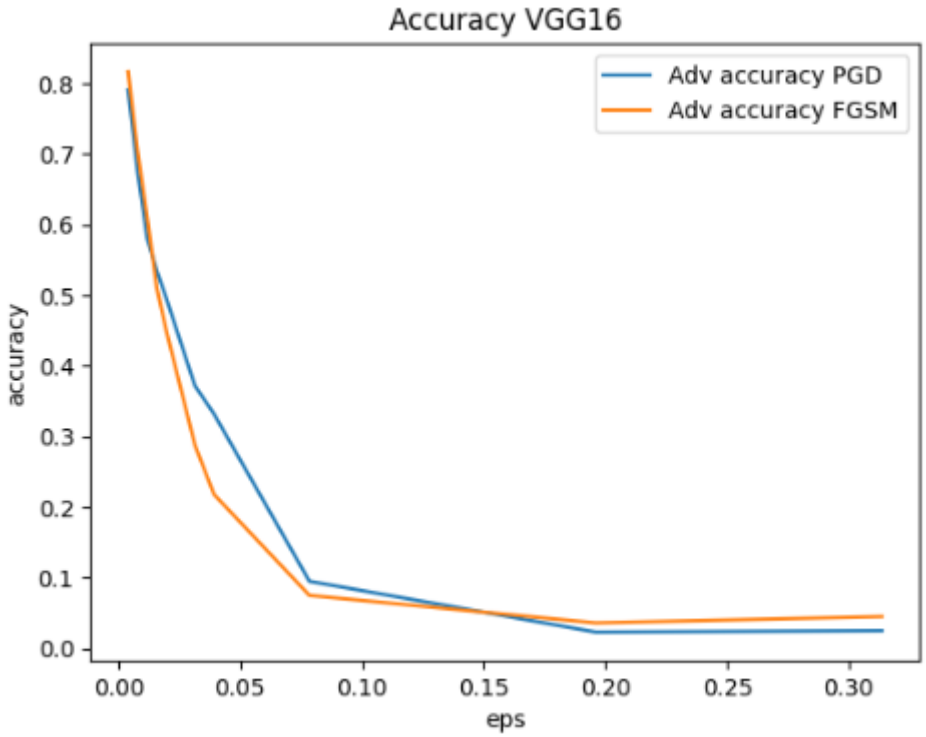
Задание 2

Для второго задания использовали тысячу первых тестовых изображений.

ResNet50: График зависимости точности классификации от параметра искажения.



VGG16: График зависимости точности классификации от параметра искажения.

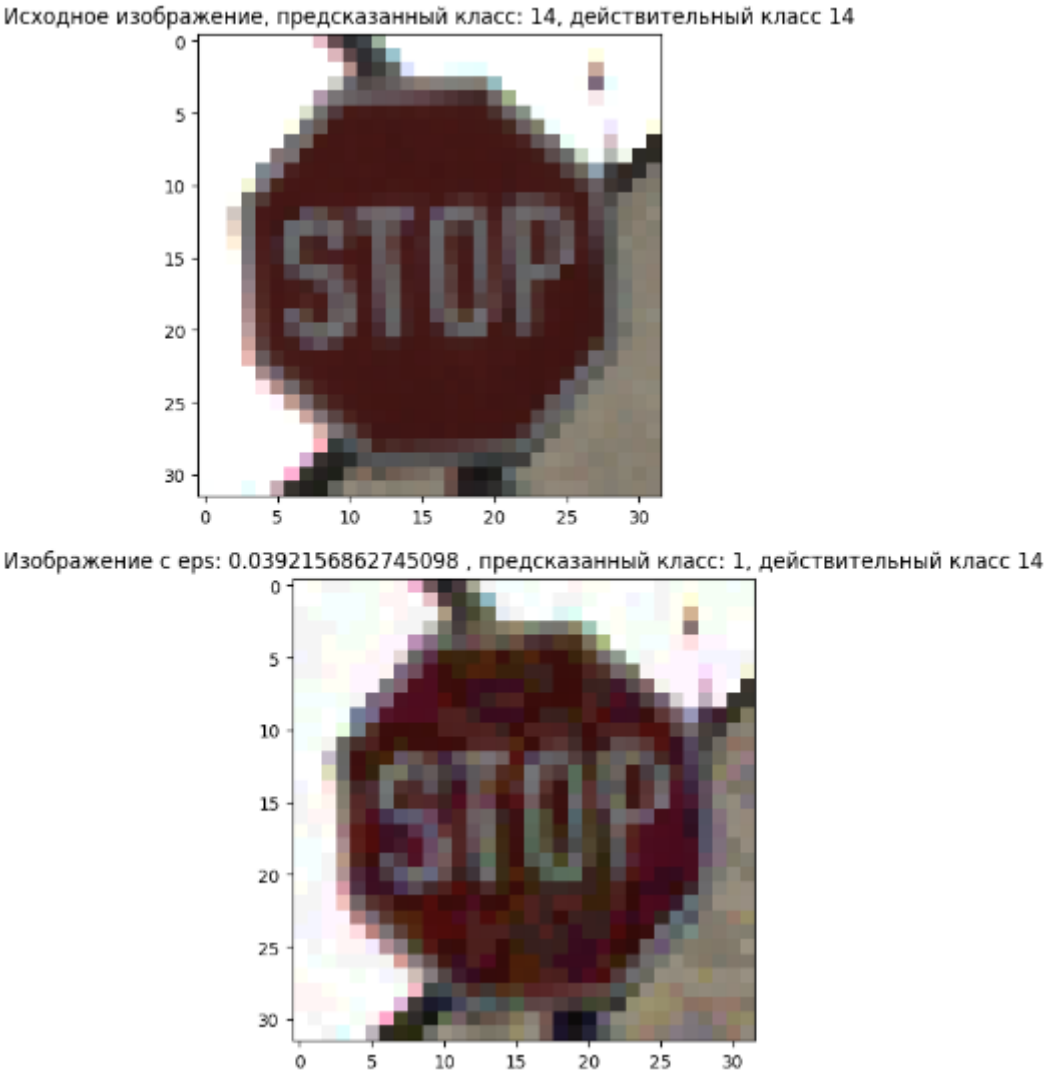


В результате была получена следующая результирующая таблица:

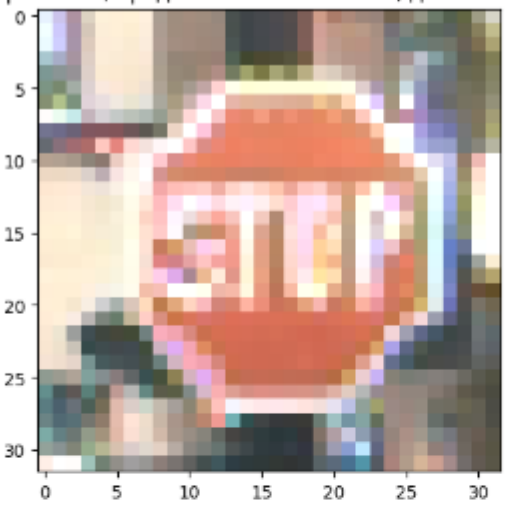
Модель	Исходные изображения	Adversarial images ϵ=1/255	Adversarial images ϵ=5/255	Adversarial images ϵ=10/255
VGG16-FGSM	92%	82%	45%	22%
VGG16-PGD	92%	79%	50%	33%
ResNet50-FGSM	92%	77%	40%	19%
ResNet50-PGD	92%	75%	37%	24%

Задание 3

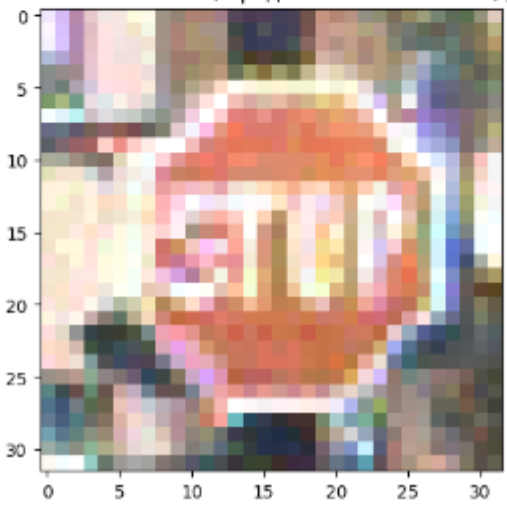
FGSM: Пример исходных изображений знака «Стоп» и соответствующих атакующих примеров.



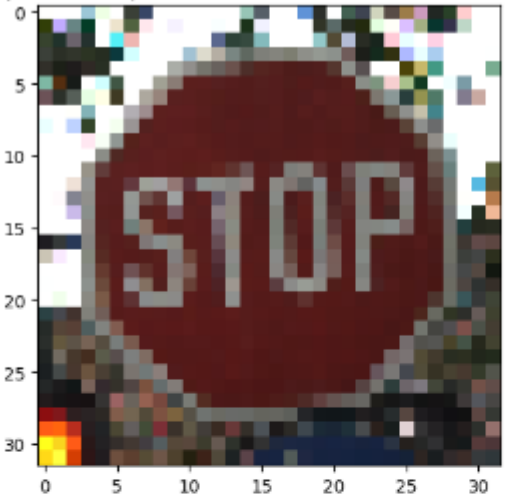
Исходное изображение, предсказанный класс: 14, действительный класс 14



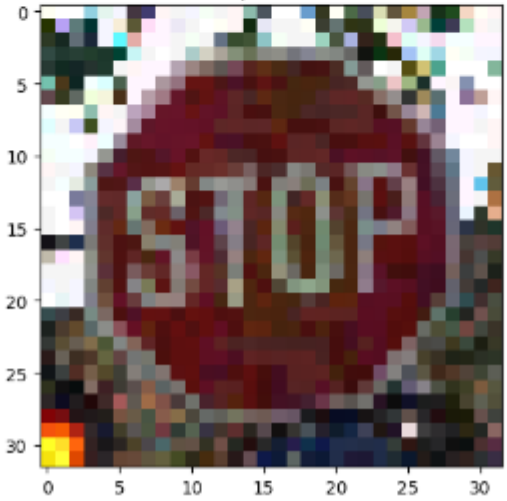
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



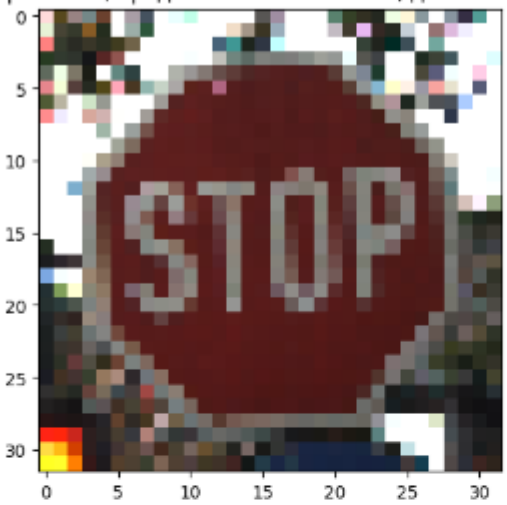
Исходное изображение, предсказанный класс: 14, действительный класс 14



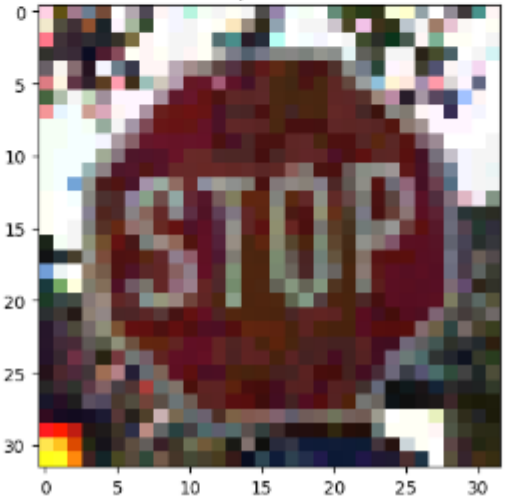
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14

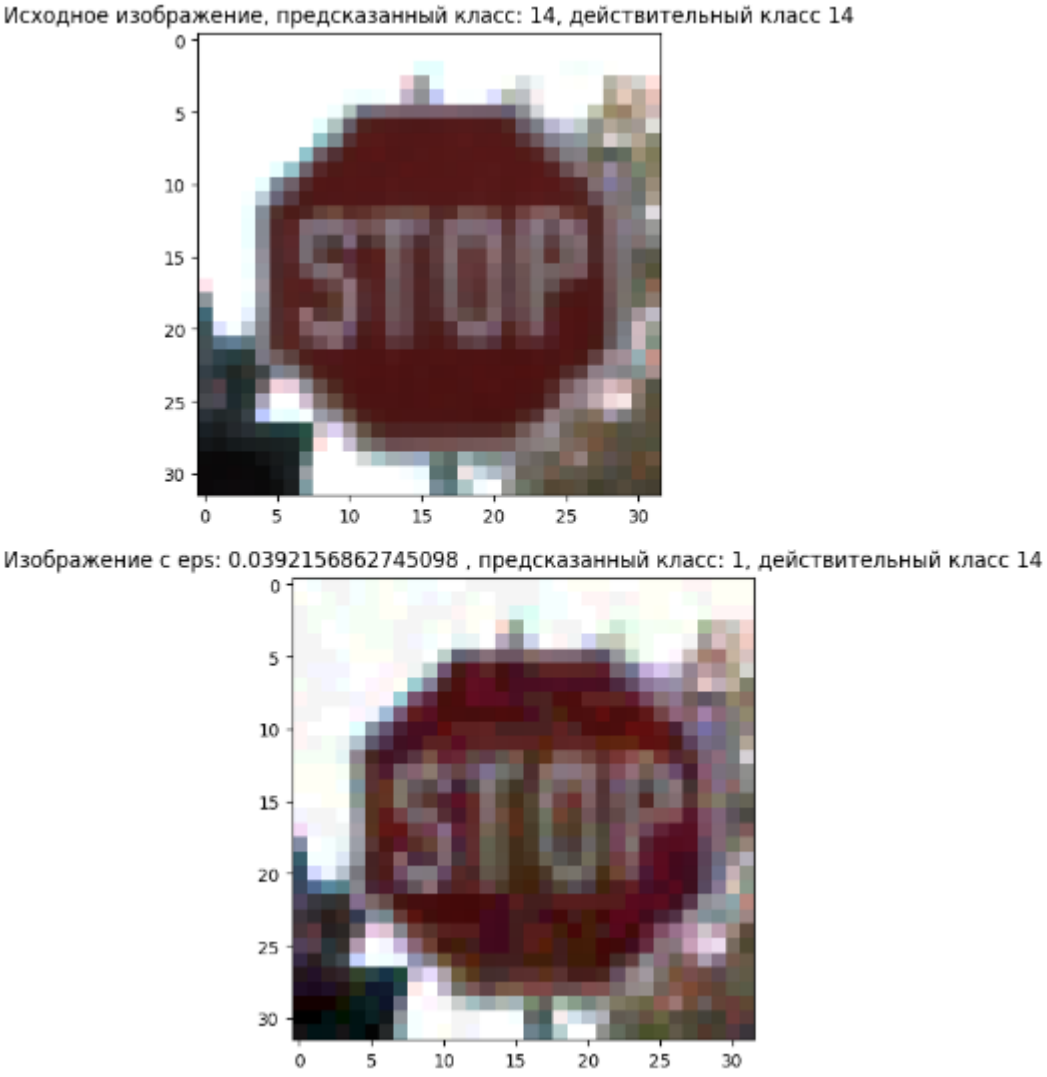


Исходное изображение, предсказанный класс: 14, действительный класс 14



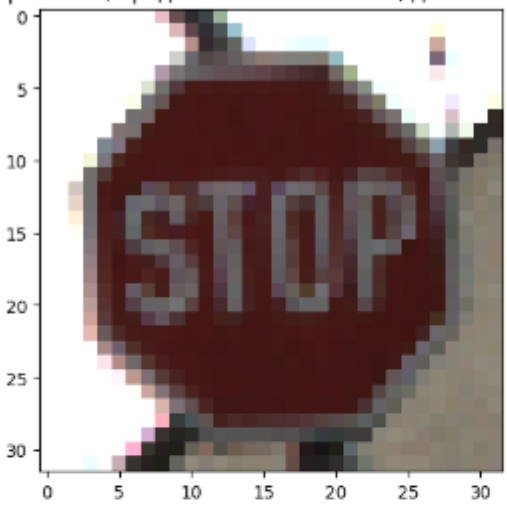
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



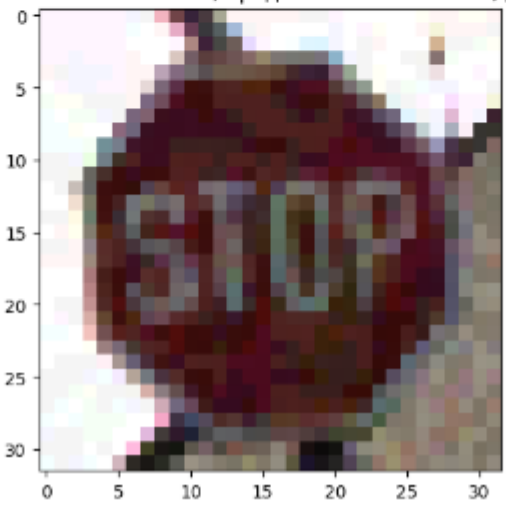


PGD: Пример исходных изображений знака «Стоп» и соответствующих атакующих примеров.

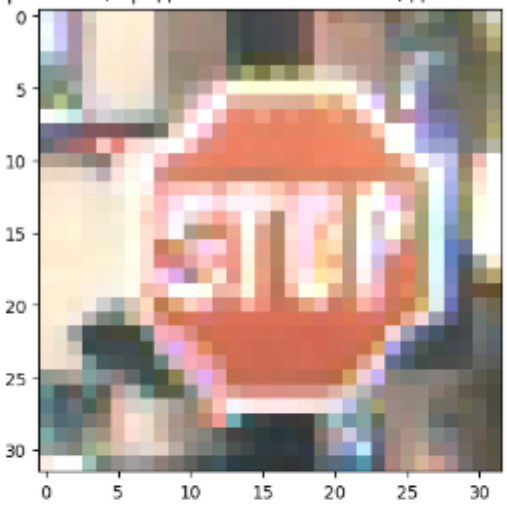
Исходное изображение, предсказанный класс: 14, действительный класс 14



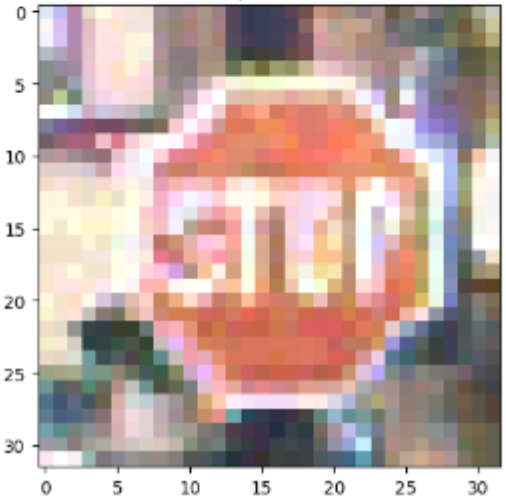
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



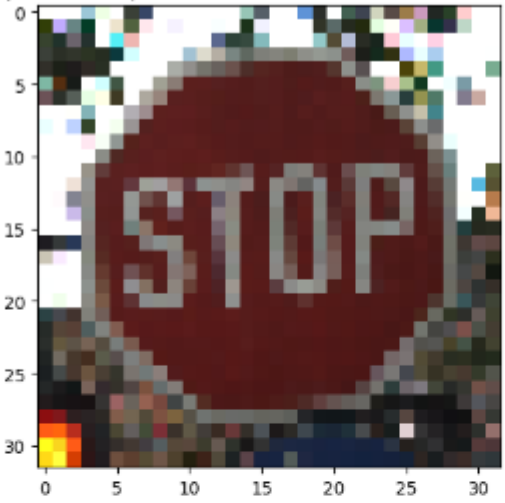
Исходное изображение, предсказанный класс: 14, действительный класс 14



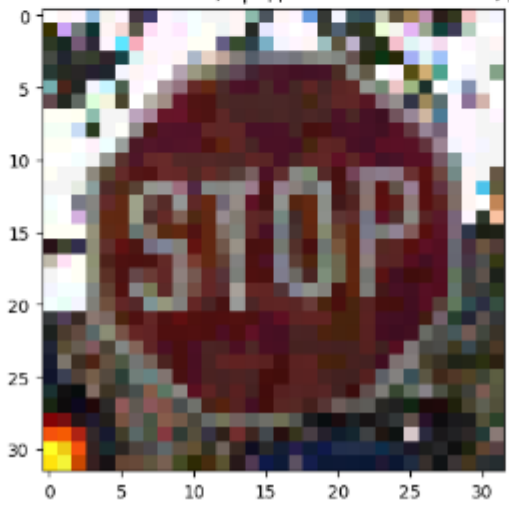
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14



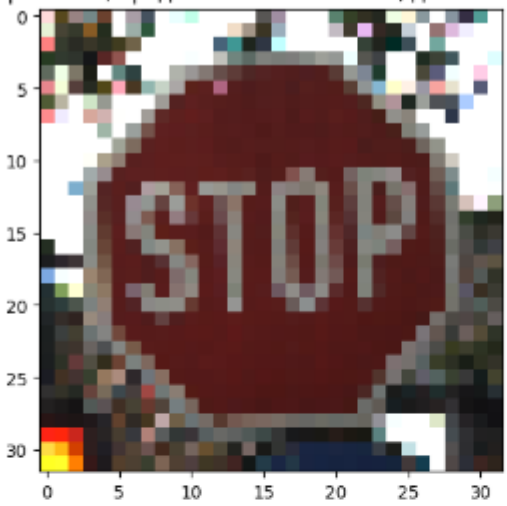
Исходное изображение, предсказанный класс: 14, действительный класс 14



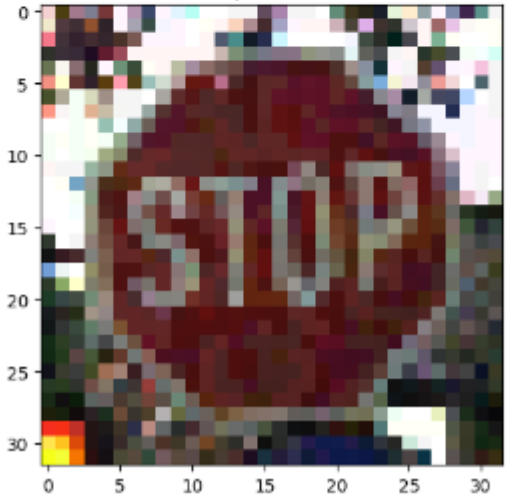
Изображение с eps: 0.0392156862745098 , предсказанный класс: 1, действительный класс 14

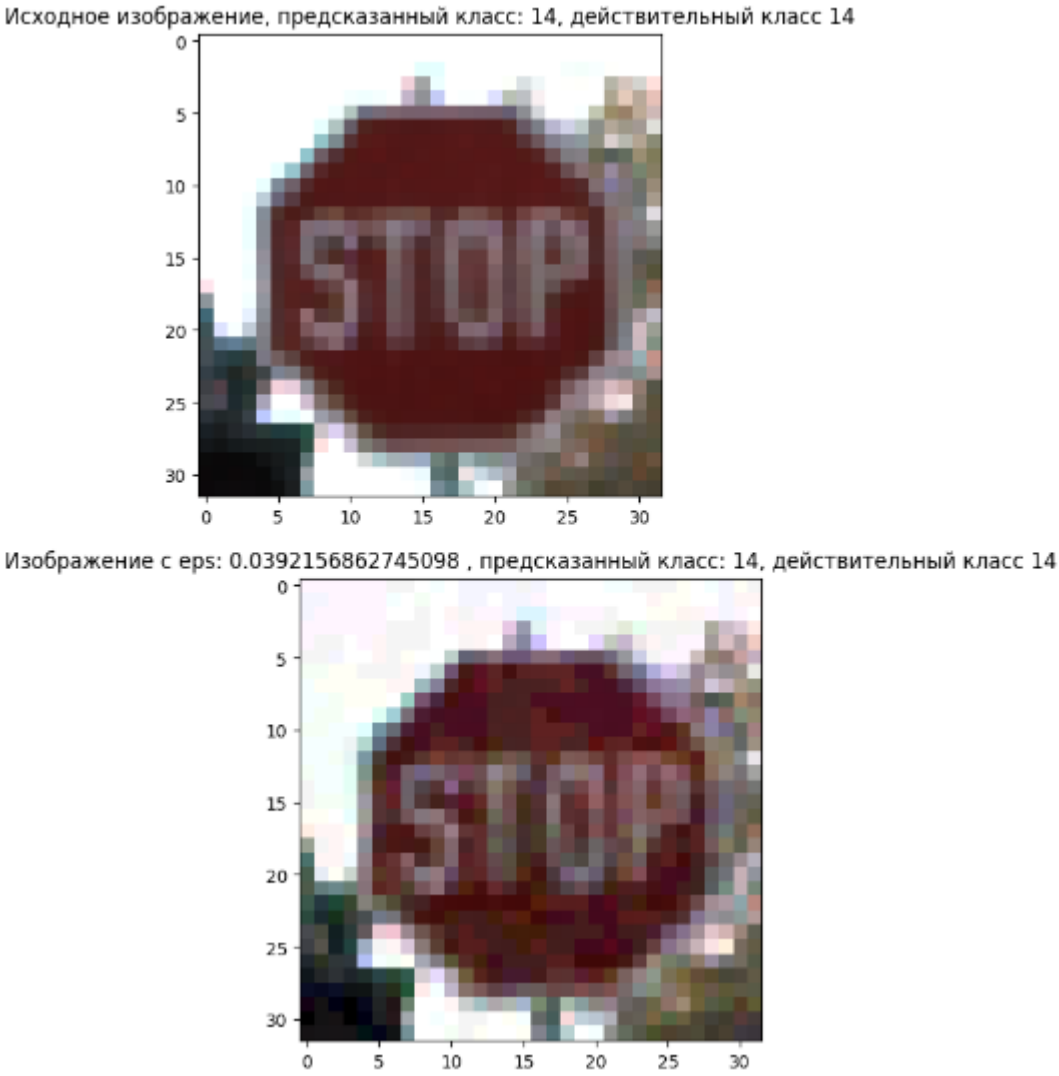


Исходное изображение, предсказанный класс: 14, действительный класс 14



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	98%	93%
ϵ=3/255	92%	76%
ϵ=5/255	80%	59%
ϵ=10/255	80%	11%
ϵ=20/255	34%	0%
ϵ=50/255	2%	0%
ϵ=80/255	1%	0%

Вывод: Метод Fast Gradient Sign Method (FGSM) неэффективен для выполнения целевых атак, поскольку при увеличении искажения, модель начинает выдавать ошибки в классификации. Вместо этого, для целевых атак более предпочтительным является метод Projected Gradient Descent (PGD). Даже если искажение сильно велико, модель все равно будет определять класс, который мы задали.