|  |
| --- |
|  |
| 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 №3

по дисциплине: «Анализ защищенности систем искусственного интеллекта»

**Выполнил**:

Студент группы ББМО-01-22

Загороднов Егор Алексеевич

**Проверил**:

К.т.н. Спирин Андрей Андреевич

Москва 2023

# **Ход выполнения работы**

В начале выполняем установку библиотеки **tf-keras-vis** для визуализации и интерпретации моделей глубокого обучения. Она позволяет проводить исследования и анализ моделей и их результатов с использованием различных методов визуализации и интерпретации данных. Установка осуществляется с помощью команды "**pip install tf-keras-vis**".

Установка **tf-keras-vis**. представлена на рисунке 1.

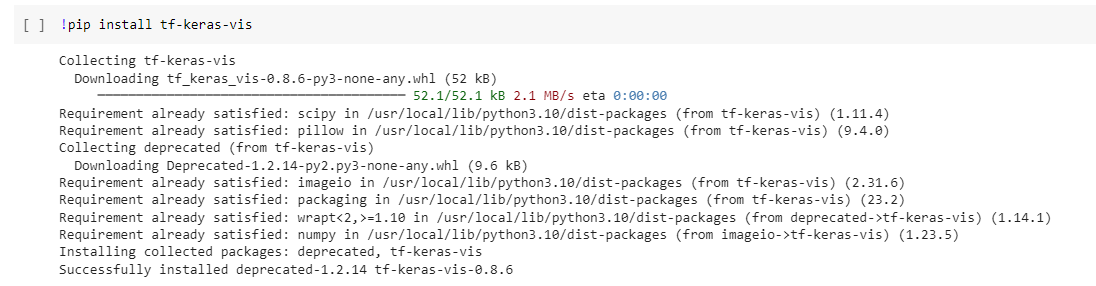


Рисунок 1 – Установка tf-keras-vis

После того как установили tf-keras-vis приступаем к загрузке модели VGG16. Модель VGG16 является одной из самых популярных моделей для обработки изображений. Она обладает высокой точностью и способна распознавать большое количество объектов.

Загрузка данной модели показана на рисунке 2.

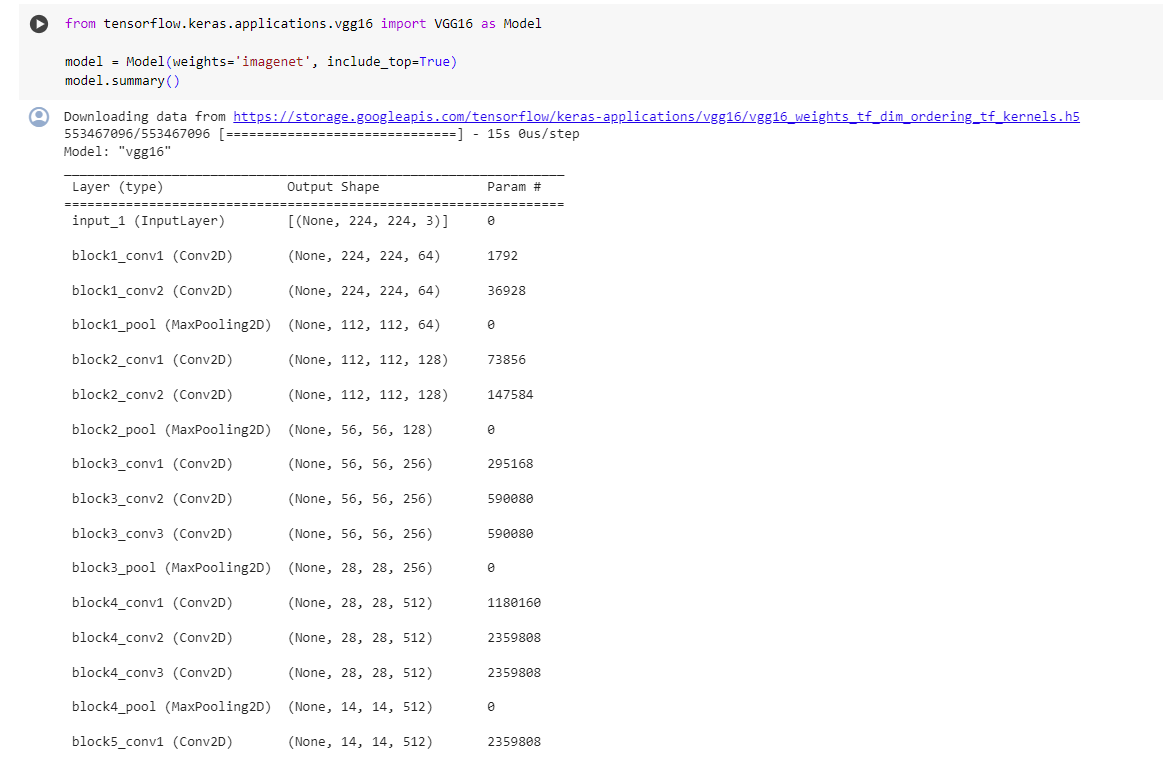


Рисунок 2 –Загрузка модели VGG16

После того, как мы успешно погрузили модель, приступаем к процессу загрузки и предобработке изображений. В данном случае, tf-keras-vis поддерживает визуализацию нескольких изображений одновременно, поэтому мы выбрали четыре различных изображения для загрузки и обработки: акулу, петуха, кукурузу и туалетную бумагу.

Загрузка и сами изображения наших объектов показаны на рисунках 3 и 4.



Рисунок 3 – Загрузка изображений

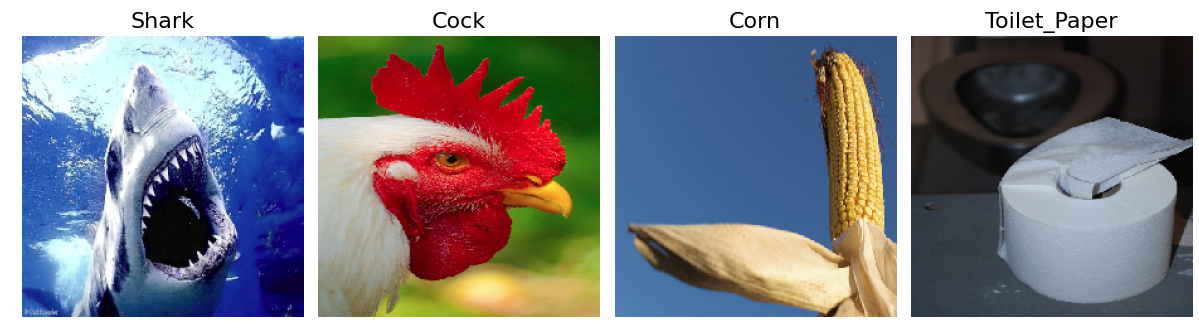


Рисунок 4 – Визуализация изображений

Приступаем к реализации функций, необходимых для использования внимания.

1) **Модификатор модели**.

Если вместо функции активации softmax мы используем функцию линейной активации в последнем слое модели, это поможет улучшить способность модели создавать точки внимания на изображениях. Для замены функции активации мы можем создать экземпляр класса **ReplaceToLinear**, который будет выполнять эту задачу. Однако, помимо этого, нам также доступна возможность использования функции модификатора модели, которую мы определили самостоятельно. Это позволяет нам более гибко контролировать и настраивать процесс модификации модели для достижения оптимальных результатов. Использование класса **ReplaceToLinear** показано на рисунке 5.

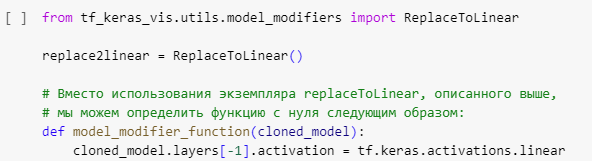


Рисунок 5 – Использование класса ReplaceToLinear

2) **Функция расчета**.

Как альтернативу функции активации softmax, мы можем использовать функцию линейной активации в последнем слое модели, чтобы избежать возможных проблем с созданием изображений внимания. Для этого мы создаем экземпляр ReplaceToLinear, который позволяет нам заменить функцию активации на линейную. Кроме использования ReplaceToLinear, мы также можем использовать функцию модификатора модели, которую мы определили сами, чтобы изменять score function. В данном случае, score function возвращает количество очков, соответствующее каждому из элементов - акуле, петуху, кукурузе и туалетной бумаге.

Процесс показан на рисунке 6.

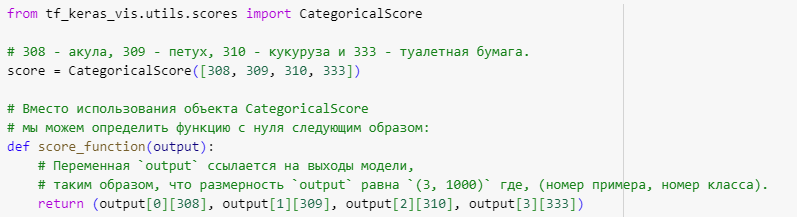


Рисунок 6 – Использование score function

**3) Ванильное внимание.**

Saliency используется для создания карты внимания, которая выделяет наиболее важные области входного изображения, которые влияют на окончательный результат. Карта визуализирует области на изображении, которые имеют наибольшее влияние на выходные значения алгоритма. Карта внимания может помочь нам понять, какие части изображения являются ключевыми для получения нужного результата. Создание объекта внимание и визуализация изображений показаны на рисунке 7.

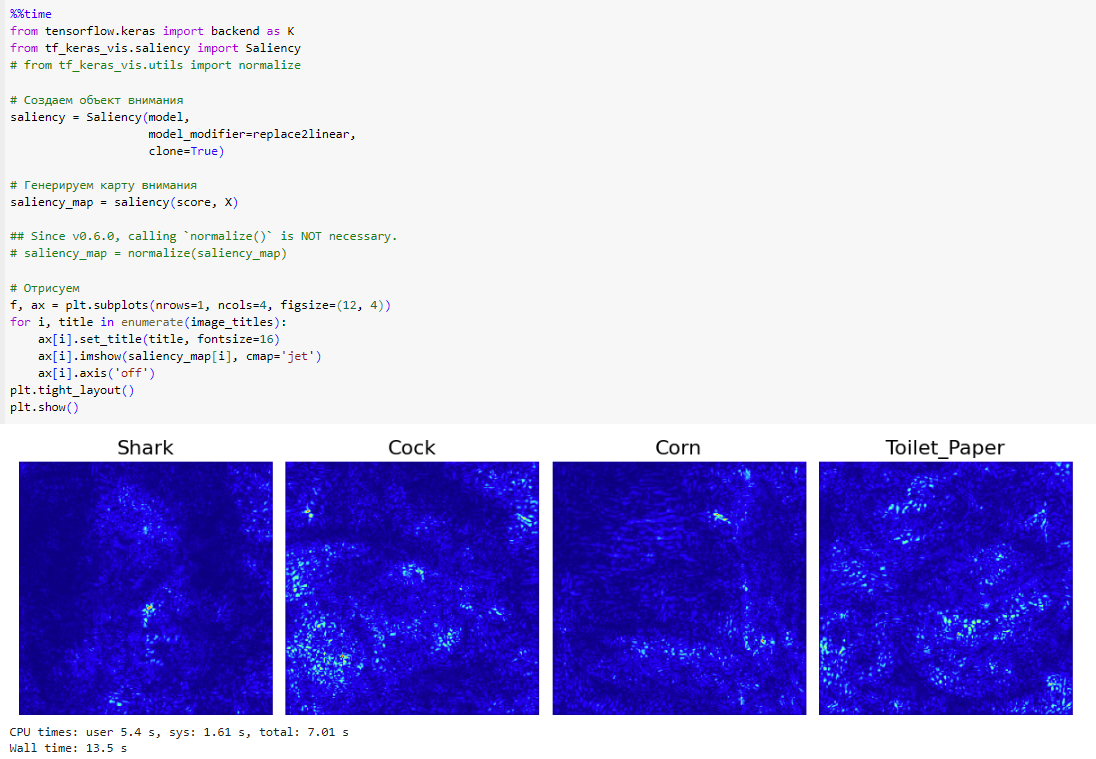


Рисунок 7 – Создание объекта внимание и его визуализация

**4) SmoothGrad.**

Генерация карты внимания, при которой выделяются области на входном изображении с наибольшим воздействием на выходное значение, предоставляет возможность выделить важные элементы.  
Для улучшения качества карты значимости можно применить метод SmoothGrad, который позволяет уменьшить шум на карте, добавляя шум к входному изображению. Это поможет получить более точное представление о влиянии каждой области на итоговый результат. Генерирование карты внимание и визуализация объектов показаны на рисунке 8.

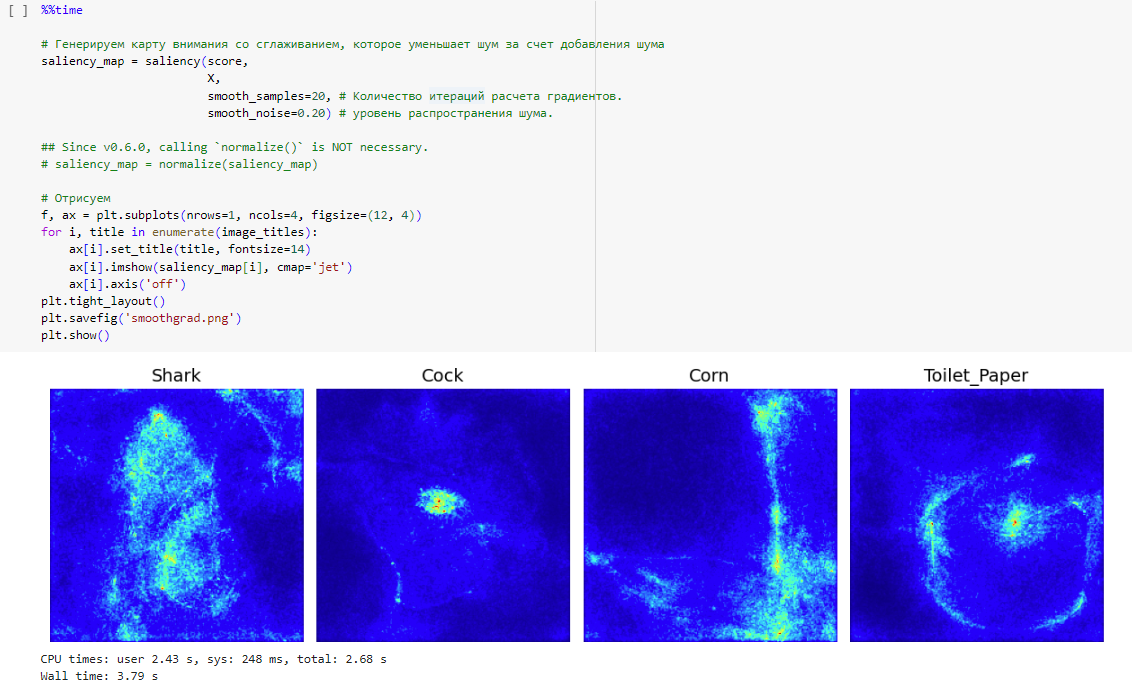


Рисунок 8 – Генерирование объекта внимания и его визуализация

**5) GradCam.**

GradCAM представляет собой альтернативный способ визуализации внимания к входным данным. В отличие от использования градиентов вывода модели, он использует данные вывода предпоследнего слоя - сверточного слоя, находящегося непосредственно перед плотными слоями. Этот подход позволяет получить более детальную информацию о том, какие области изображения привлекли наибольшее внимание модели. Как показано ниже на рисунке 9, метод GradCAM предоставляет полезную интуитивную информацию о том, на чем сфокусировано внимание. Тем не менее, при более внимательном рассмотрении видно, что отображаемое внимание не охватывает всю цель на изображении.

Создание объекта GradCam, генерирование тепловой карты с GradCam и визуализация изображений показаны на рисунке 9.

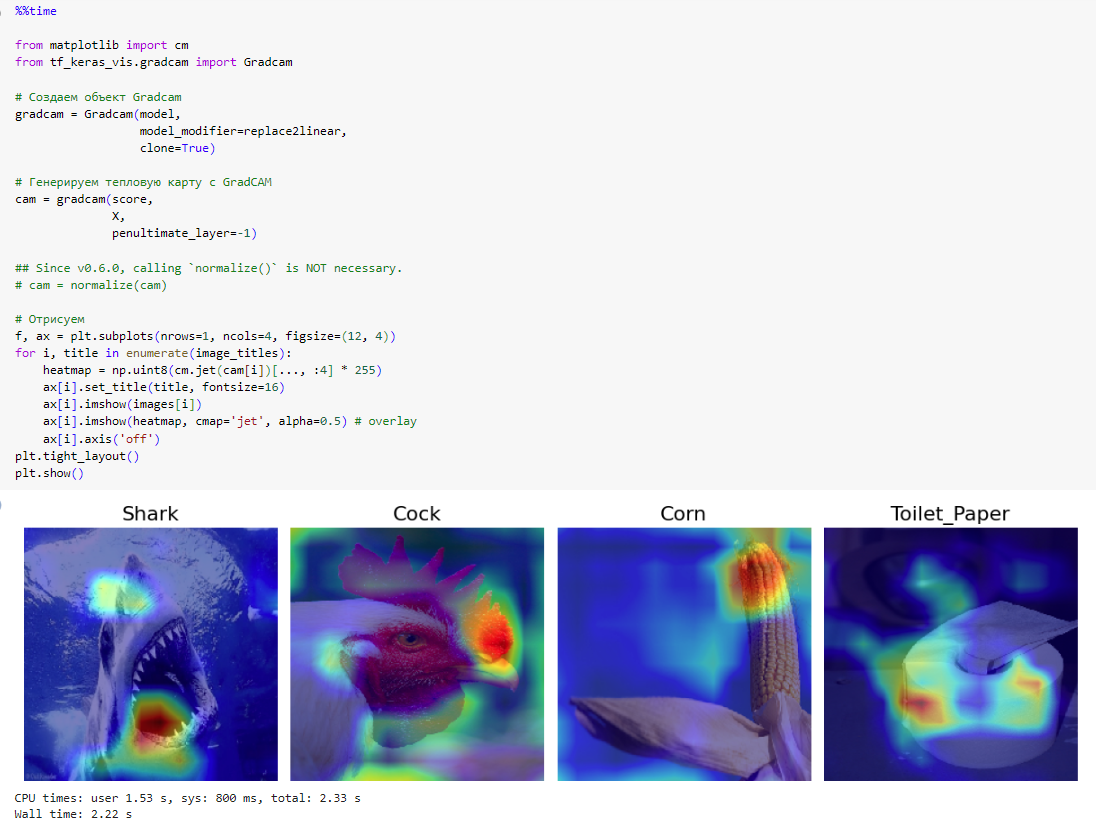


Рисунок 9 – Создание объекта Gradcam, генерирование тепловой карты и визуализация изображений

**6) GradCam++**

GradCAM++ может обеспечить лучшее визуальное объяснение прогнозов модели, чем GradCAM. Он предоставляет более точное отображение внимания на объекты на изображении, полностью покрывая цель. Этот метод основан на более сложных вычислениях и учитывает более широкий контекст изображения, что делает его более надежным и информативным визуальным объяснением прогнозов модели CNN. Таким образом, использование GradCAM++ может быть предпочтительным в ситуациях, где мы хотим получить более детальное и полное представление о том, на чем модель сосредотачивает свое внимание. Создание объекта Gradcam++, генерирование тепловой карты с GradCAM++ и визуализация изображений показаны на рисунке 10.

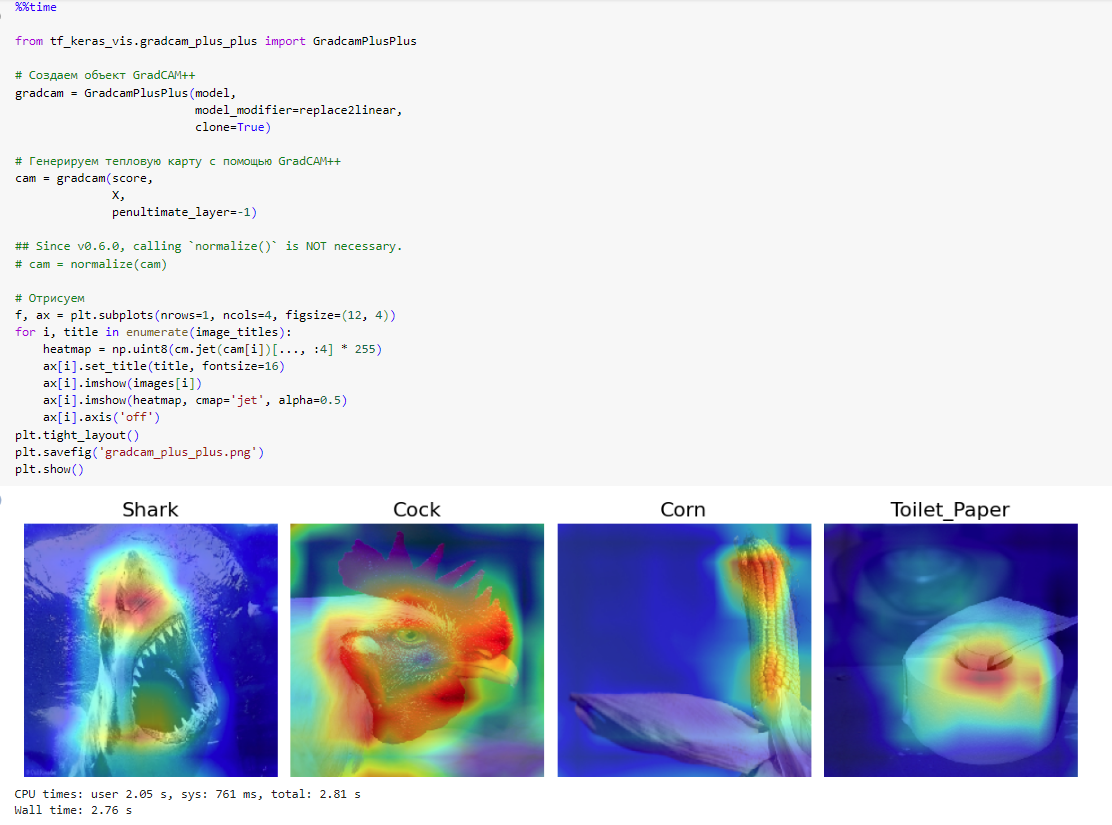


Рисунок 10 – Создание объекта Gradcam++, генерирование тепловой карты и визуализация изображений

# Заключение

В процессе выполнения лабораторной работы были реализованы указанные задачи, а именно:

- Созданы два классификатора с использованием глубоких нейронных сетей на наборе данных GTSRB;

- Применена нецелевая атака уклонения с использованием метода белого ящика на модели глубокого обучения;

- Применена целевая атака уклонения методом белого ящика на модели глубокого обучения.