|  |
| --- |
|  |
| 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-02-22

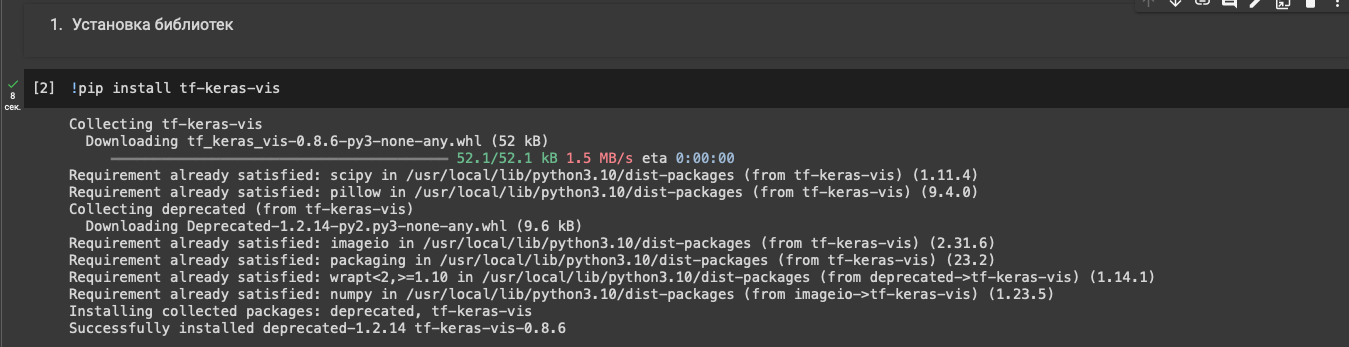
Филиппов Леонид Алексеевич

**Проверил**:

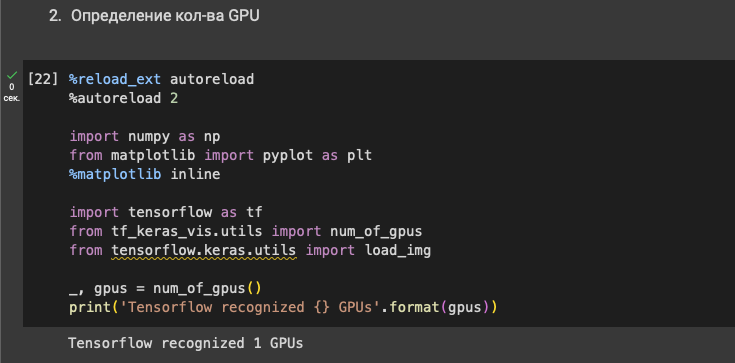
Спирин Андрей Андреевич

Москва 2024

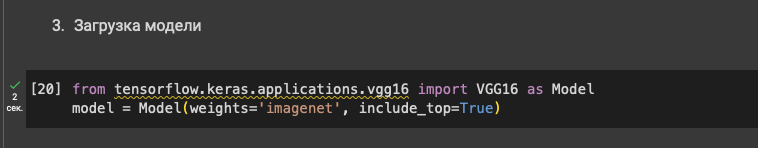
1. Установка библиотек



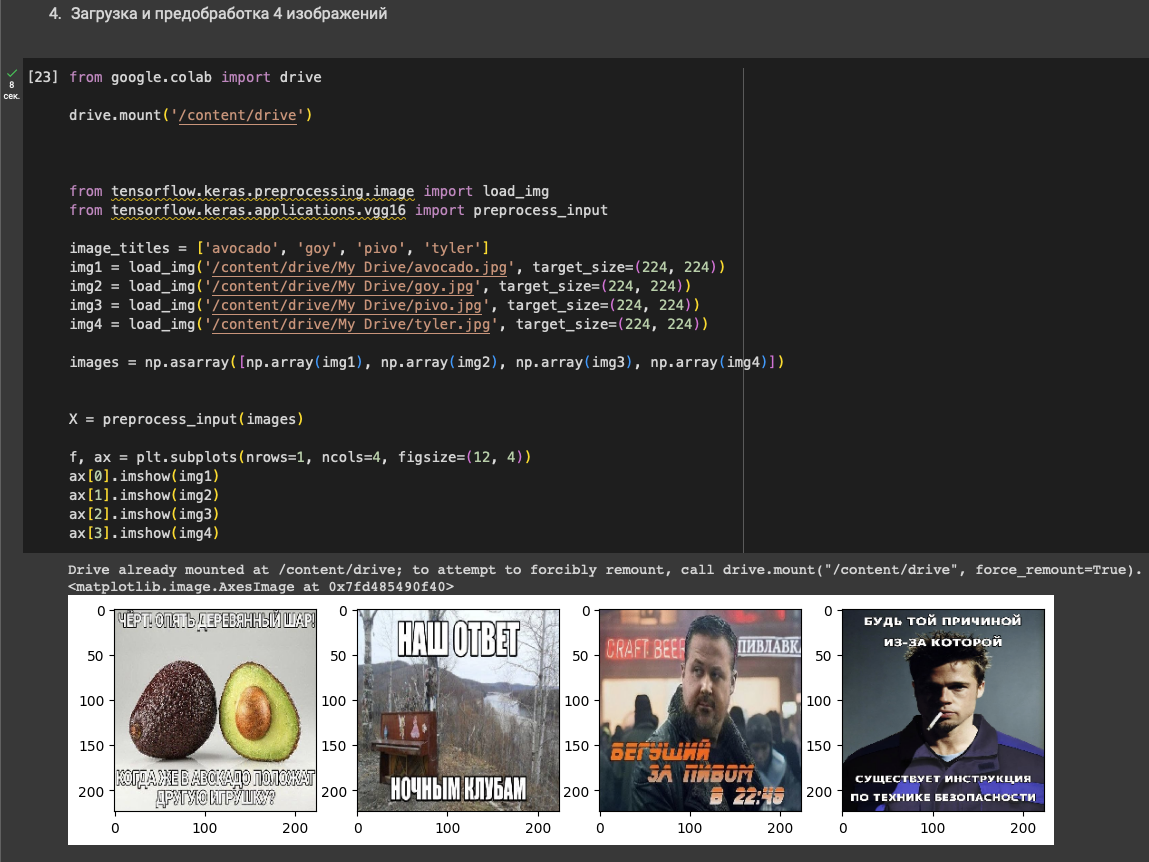
1. Определение кол-ва GPU



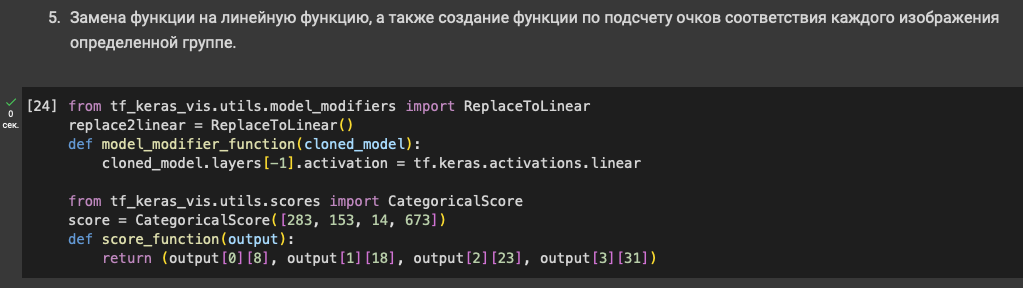
1. Загрузка модели



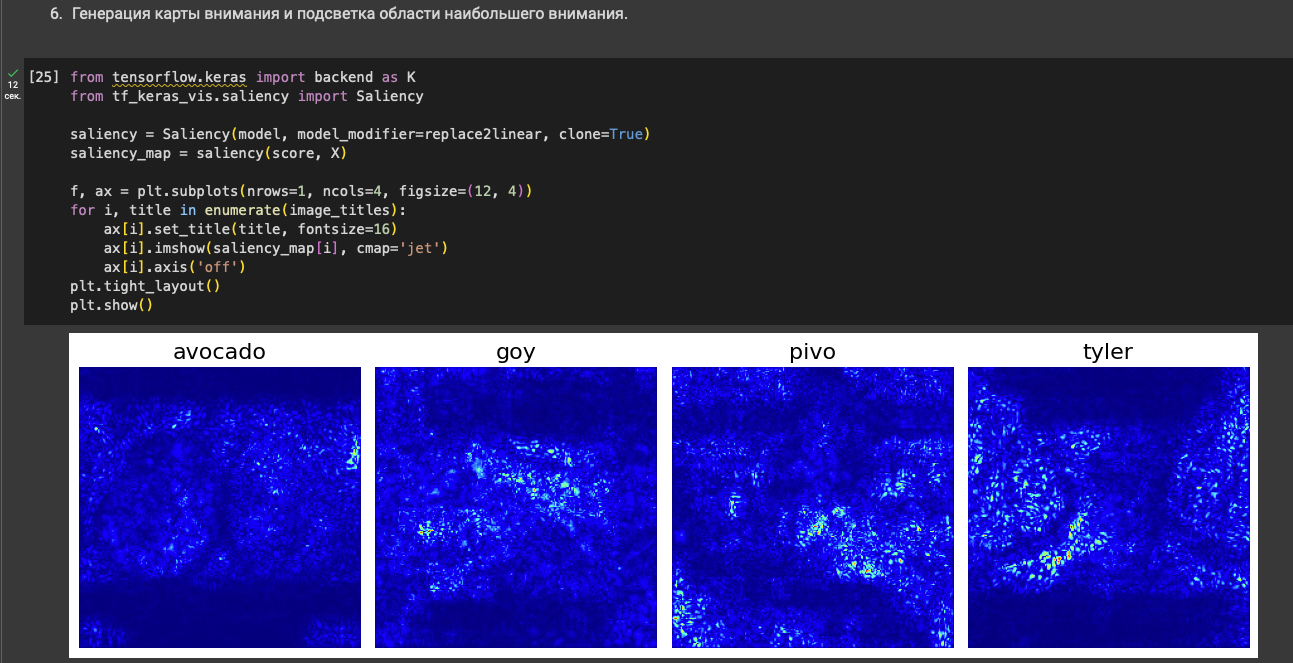
1. Загрузка и предобработка 4 изображений



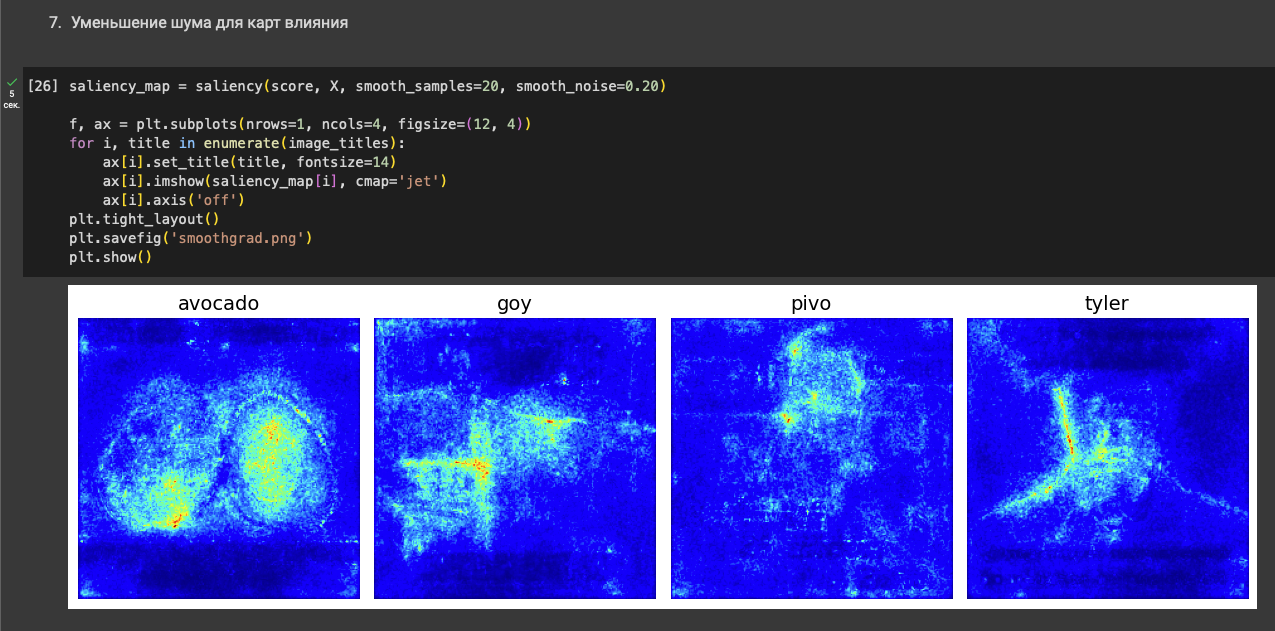
1. Замена функции на линейную функцию, а также создании функции по подсчету очков соответствия каждого изображения определенной группе



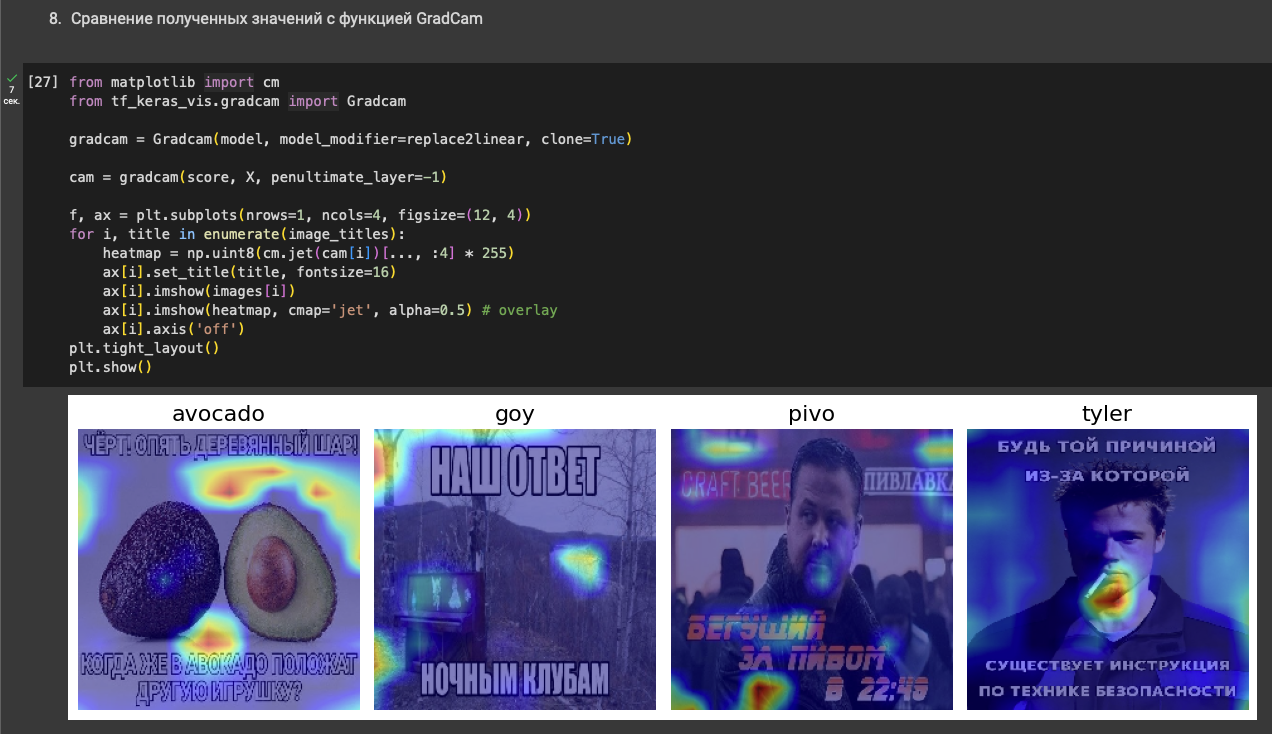
1. Генерация карты внимания и подсветка области наибольшего внимания



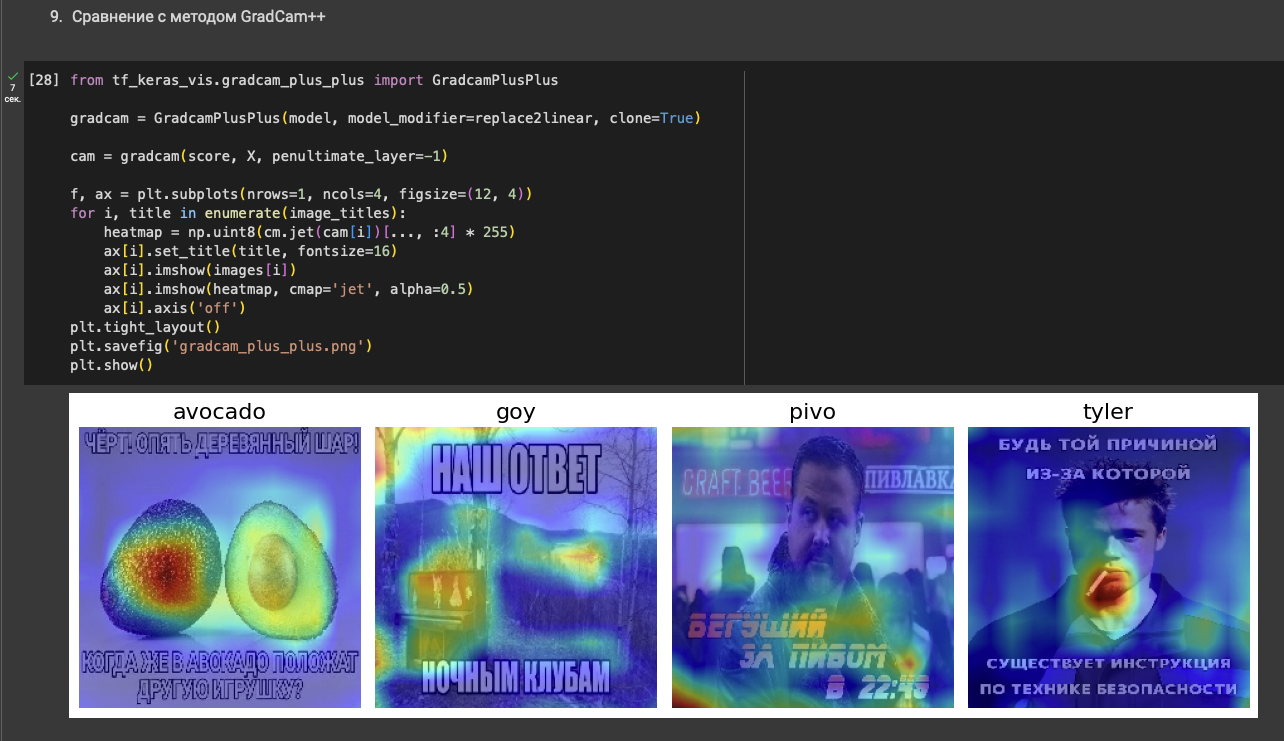
1. Уменьшение шума для карт влияния



1. Сравнение полученных значений с функцией GradCam



1. Сравнение с методом GradCam++



**Вывод**

SmoothGRAD, GradCAM и GradCAM++ представляют собой методы визуализации, предназначенные для анализа воздействия различных областей изображения на решения нейронных сетей.

SmoothGRAD использует метод усреднения градиентов с добавлением случайного шума к входным данным. Это прием помогает сгладить изображение и уменьшить воздействие шума, что делает визуализацию более интерпретируемой.

GradCAM (Gradient-weighted Class Activation Mapping) сосредотачивается на активациях конкретного класса, взвешивая градиенты, проходящие через слои сети. Это позволяет выделить важные части изображения, влияющие на принятое решение.

GradCAM++ представляет собой улучшенную версию GradCAM, которая учитывает как положительные, так и отрицательные влияния активаций при создании карт активации. Это способствует более точной локализации значимых областей на изображении.