|  |  |  |
| --- | --- | --- |
|  |  |  |
| МИНОБРНАУКИ РОССИИ | | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА– Российский технологический университет»**  **РТУ МИРЭА** | | |
| Институт кибербезопасности и цифровых технологий  Кафедра КБ-4 «Интеллектуальные системы информационной безопасности» | | |

**Отчёт по лабораторной работе № 3**

**по дисциплине «Анализ защищенности систем искусственного интеллекта»**

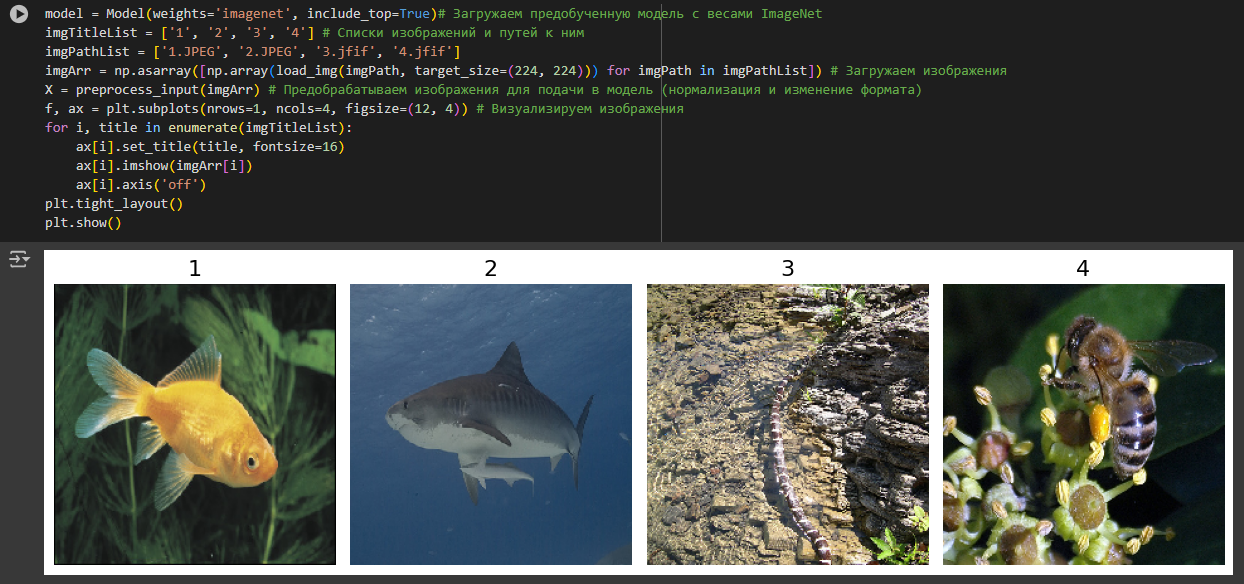
Выполнил(а)

Студент 1 курса

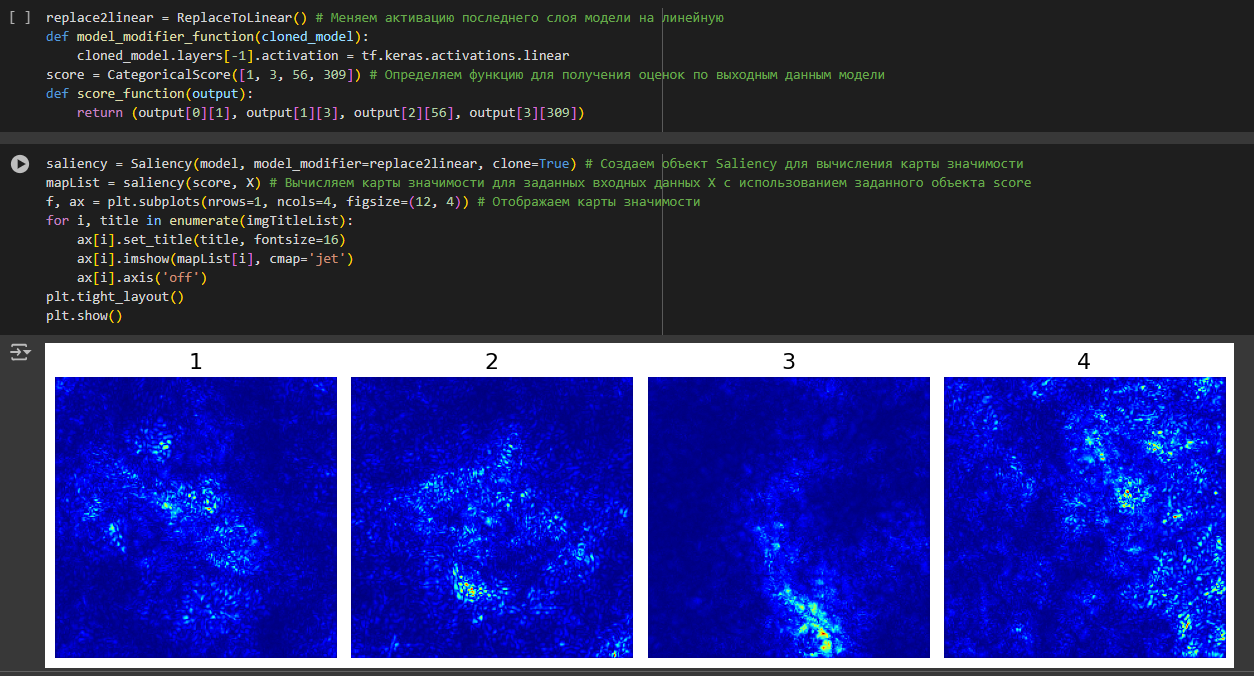
Группы ББМО-02-23

Колонин Артём Николаевич

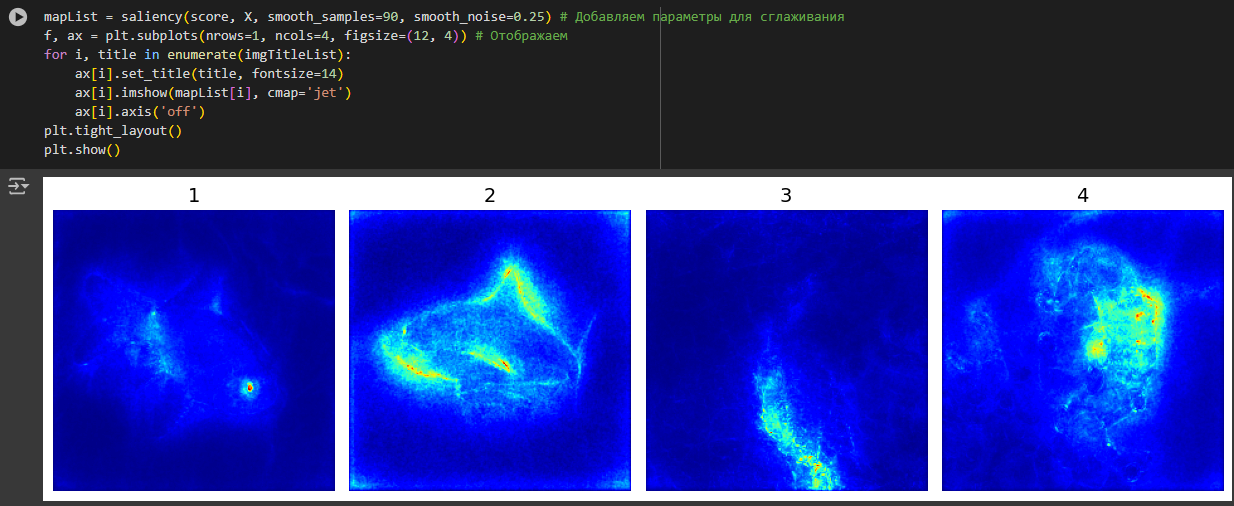
Загружаем случайные изображения из датасета imagenet. Мной выбраны изображения goldfish, tiger shark, king snake и bee



Меняем функцию активации последнего слоя на линейный, задаём метки классов выбранным изображениям (для каждого изображения соответственно метки равны 1, 3, 56, 309). Получаем карту значимости различных областей изображения.

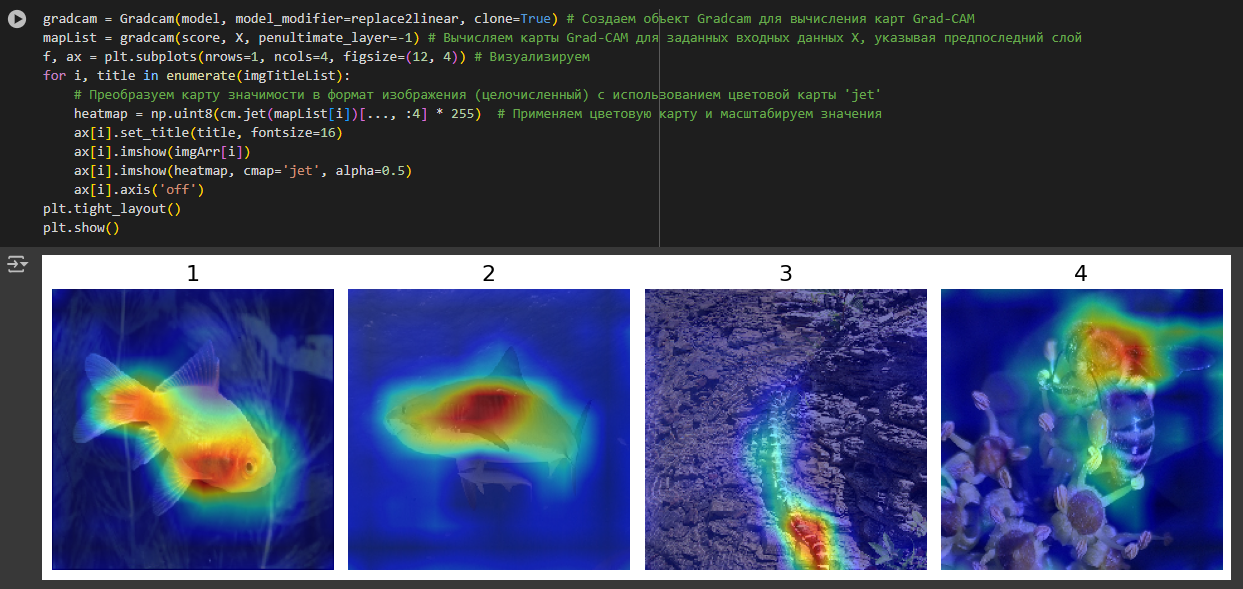


Добавим сглаживания для улучшения наглядности важности

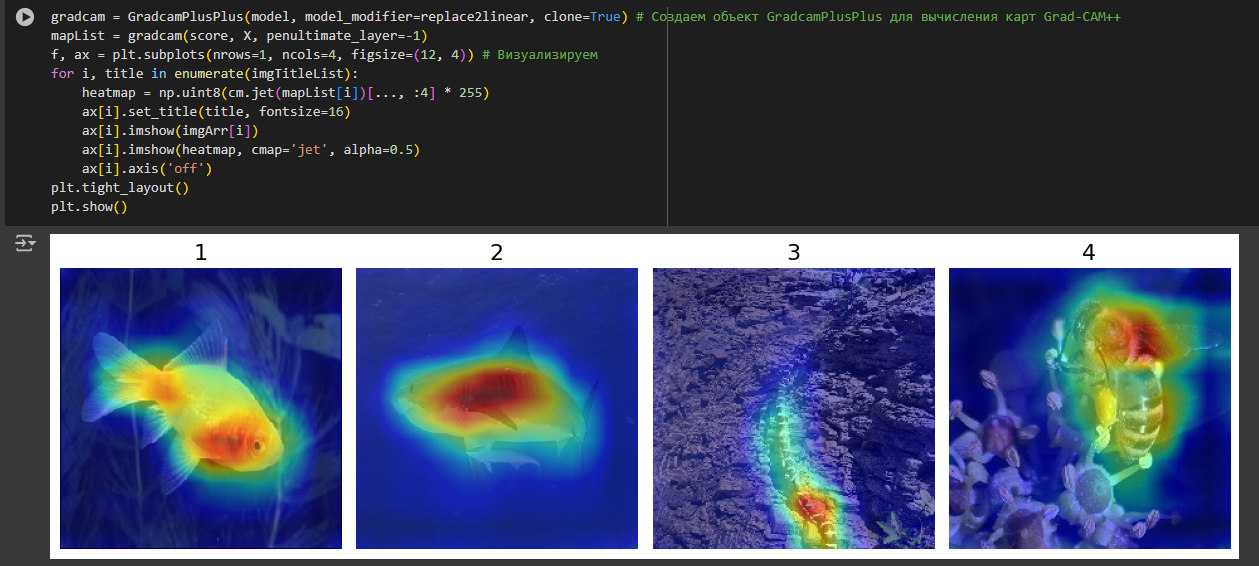


Это позволяет нам увидеть, на какие области изображения обращает внимание нейронная сеть при определении конкретного класса.

Далее используем Gradcam для более лучшей передачи важности признаков. Этот метод содержит больше пространственных деталей, которые теряются в плотных слоях.



Gradcam++, как улучшенный вариант предыдущего метода



Хотя Gradcam++ не сильно изменяет результат

В ходе данной работы были применены различные инструменты для более наглядной демонстрации работы нейронных сетей. Это может помочь в их отладке, так как помогает проверить, на какие области нейросеть обучилась обращать внимание.