|  |  |  |
| --- | --- | --- |
|  |  |  |
| 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»**

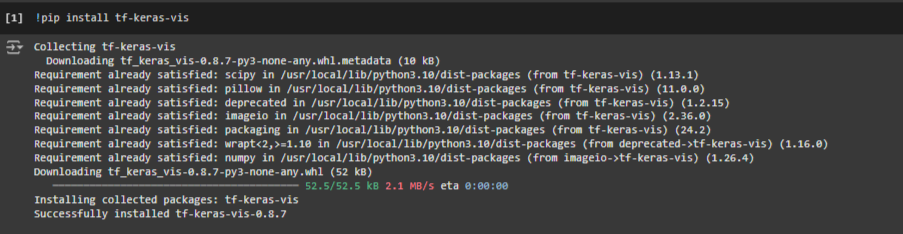
Выполнил:

Евдокимов А.М.

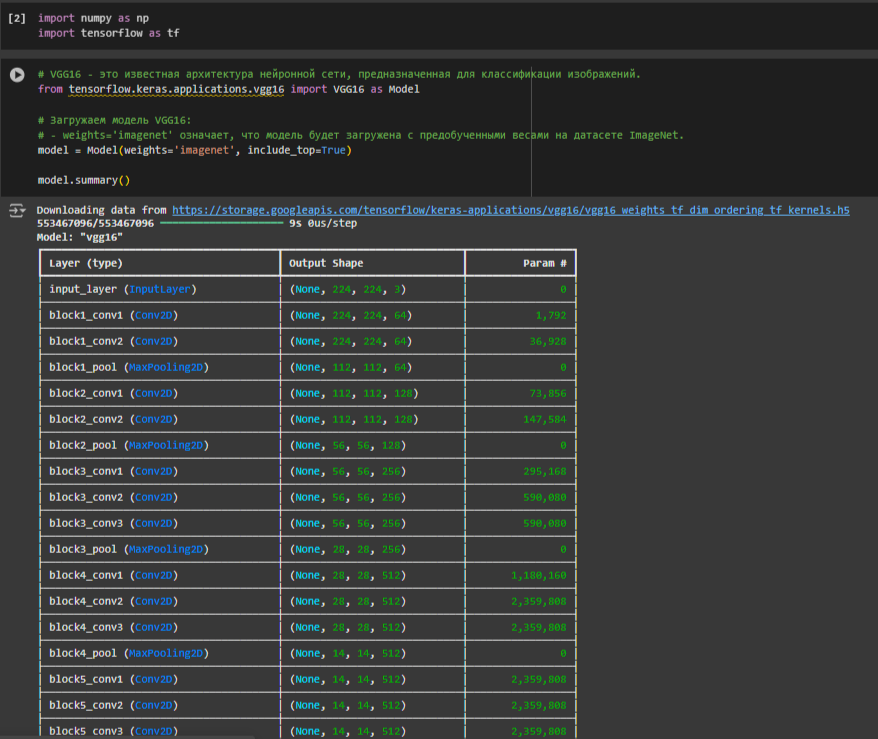
Группа: ББМО-02-23

Москва - 2024

Установим инструмент для визуализации для TensorFlow Keras.

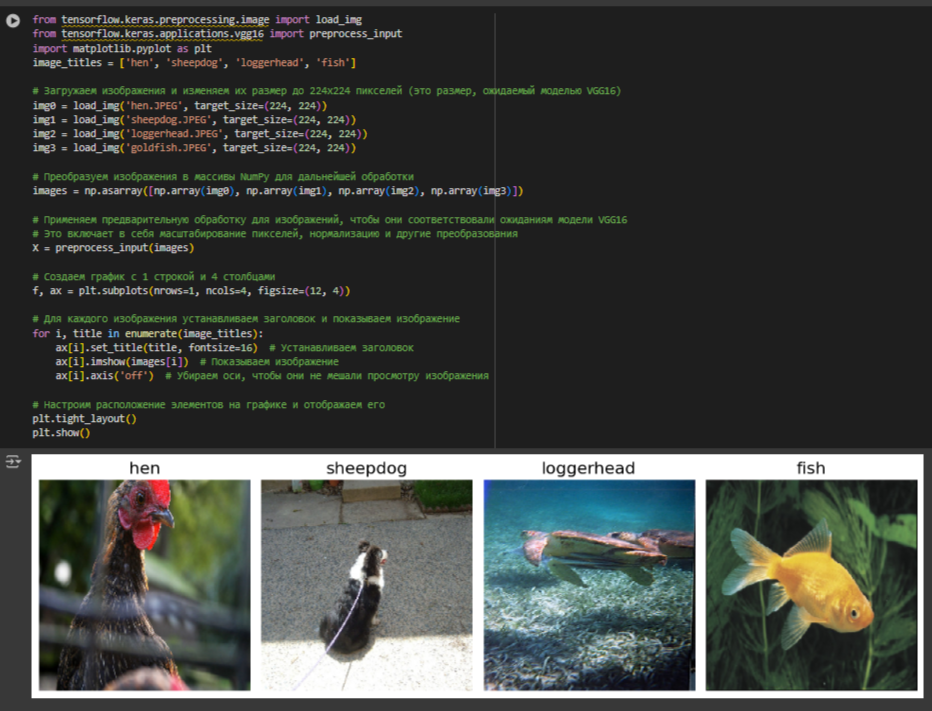


Далее подключим необходимые библиотеки. Загрузим предварительно обученную модель VGG16, на ImageNet датасете. После чего отобразим сводку по модели.

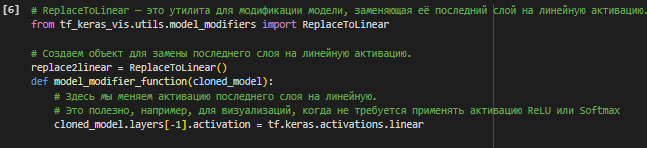


Загрузим несколько изображений датасета ImageNet и выполним их предварительную обработку перед использованием.

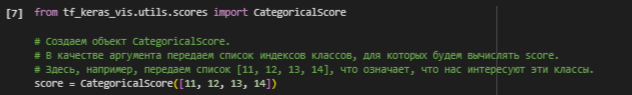
Отобразим на одном графическом представлении все наши изображения.



Заменим функцию активации на линейную



Создадим функцию очков соответствия каждому изображению



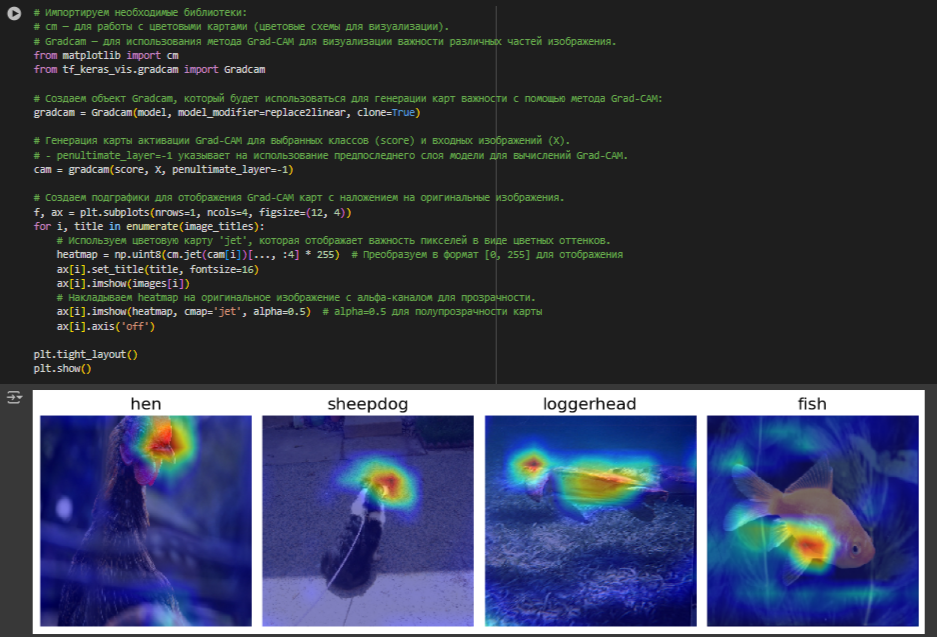
Создадим карту внимания (vanilla)



Уменьшим шум для карт влияния



Используем GrandCAM



Используем GradCAM++



Таким образом, использование методов визуализации Grad-CAM, Grad- CAM++, Saliency, SmoothGrad и подобных, может быть полезным для понимания того, какие части изображений были наиболее важными при принятии решений моделью машинного обучения. Методы Grad-CAM и Grad- CAM++ позволяют визуализировать активации в различных частях изображений, позволяя понять, где модель фокусируется при определении классов. Карты активаций могут помочь исследовать, какие объекты или части изображения были ключевыми для принятия решения моделью. Использование метода SmoothGrad может помочь снизить шум и сделать карты сайленси более интерпретируемыми.

**Вывод**

В результате выполнения работы получен некоторый опыт работы с инструментами получения карт активации для более подробного объяснения работы нейронных сетей и других моделей машинного обучения. Были также проведены эксперименты по созданию карты активации моделей машинного обучения с примерами на изображениях из набора данных ImageNet и модели VGG.

В рамках работы были рассмотрены несколько методов создания карт активации, в том числе GradCAM. Были проведены эксперименты с различными параметрами и модификациями методов.

После завершения экспериментов и проведения анализа полученных результатов были сделаны выводы касательного полноты и точности объяснения работы методов генерации карт.