|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| **Федеральное государственное бюджетное образовательное учреждение**  высшего образования  «МИРЭА – Российский технологический университет» |

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:**

ББМО–02–22

Шмарковский М. Б.

**Проверил:**

Спирин А. А.

Москва, 2024

Цель работы

Выполнить задания по построению карт активации различными способами и методами. Провести анализ полученных результатов, изменяя параметры и сделать соответствующие выводы по проделанной работе.

Задание

Настроить и провести эксперимент по генерации карт значимости признаков на наборе данных imagenet;

оценить результаты, провести дополнительные эксперименты, изменяя параметр backprop\_modifier;

выполнить построение карт значимости классов для выбранных изображений методом для интерпретирования моделей gradCAM;

cделать выводы о наиболее точном и полном методе описания активаций слоев нейронной сети.

Ход выполнения работы

В качестве набора данных для данной работы выступают 4 изображения, полученные из тестового набора данных для ознакомления. Выбраны были 4 следующих изображения (рисунок 1).



Рисунок – 1

Добавим требующиеся библиотеки и установим keras, процесс показан на рисунке 2. А также модель VGG16, показана на рисунках 3-4.

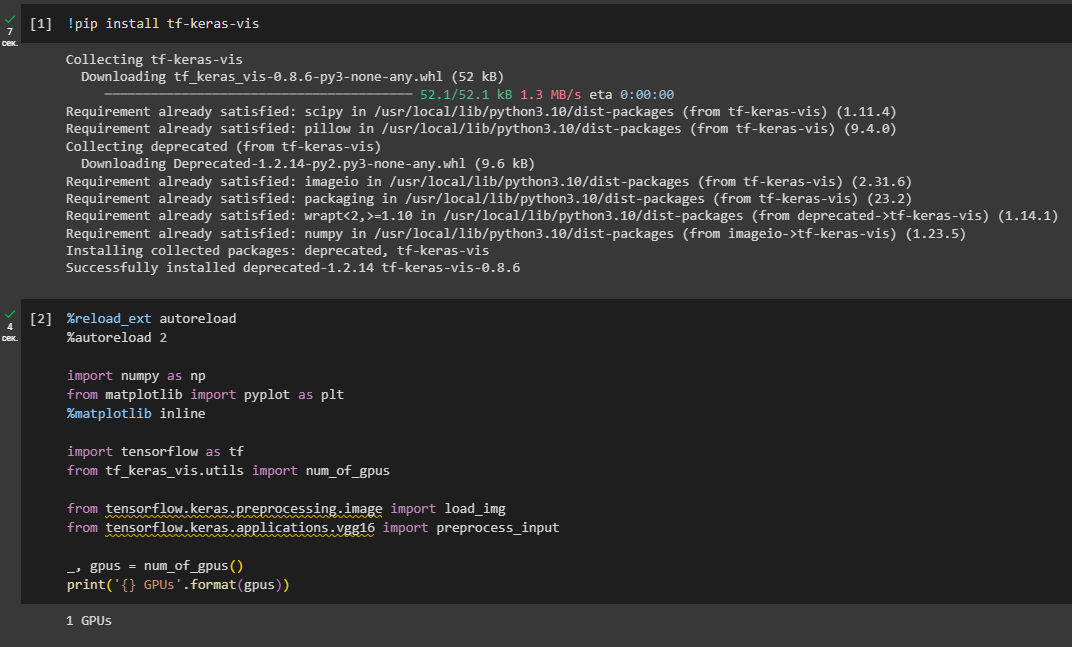


Рисунок – 2

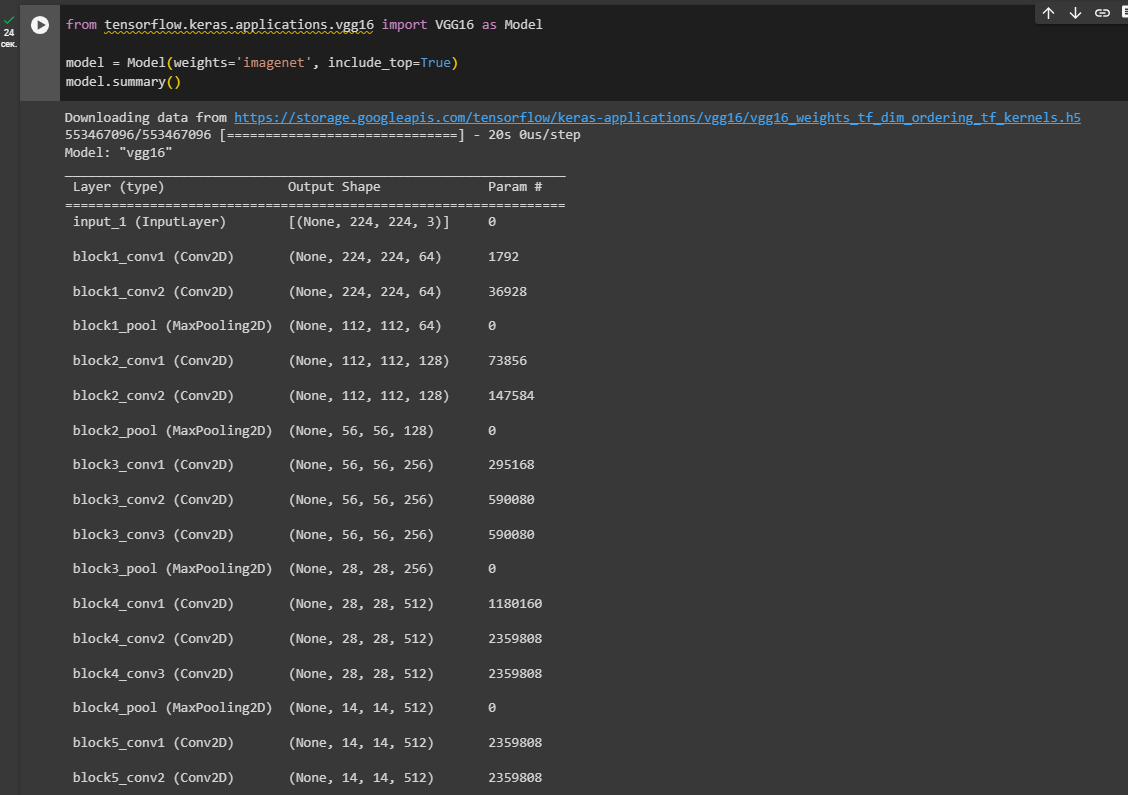


Рисунок – 3

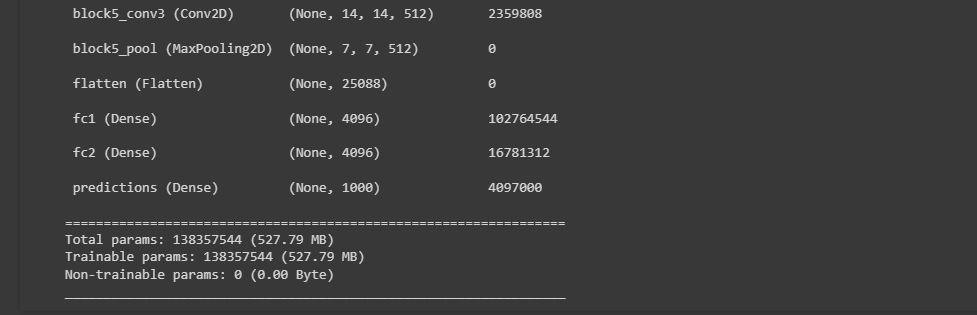


Рисунок – 4

Загрузим и подготовим изображения на рисунке 5.



Рисунок – 5

Реализуем функцию для линейной активации в последнем слое модели вместо softmax (улучшение созданий изображений внимания). А также функцию расчета score, в нашем случае 284 для кота. Показаны на рисунке 6.

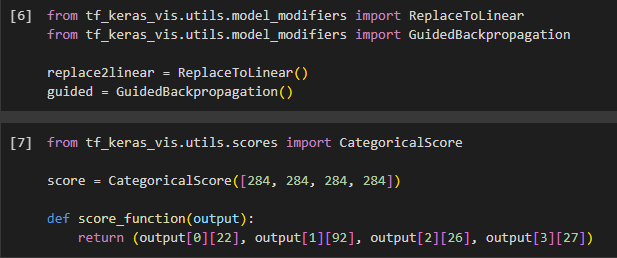


Рисунок – 6

Смотрим карты ванильного внимания. Видим низкое качество карты, коты уже вырисовываются, но пока слабо различимы. Показаны на рисунке 7.

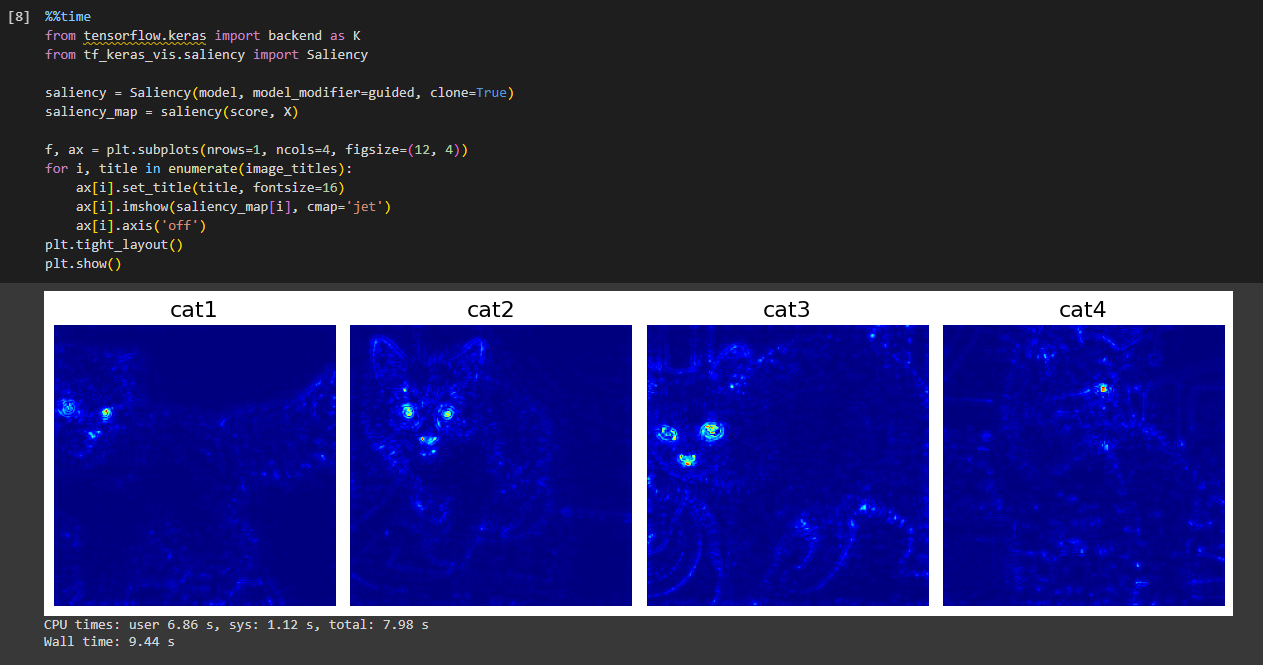


Рисунок – 7

Смотрим карты smoothgrad. Видим улучшенное качество карты, можно понять, что изначальный объект кот. Показаны на рисунке 8.

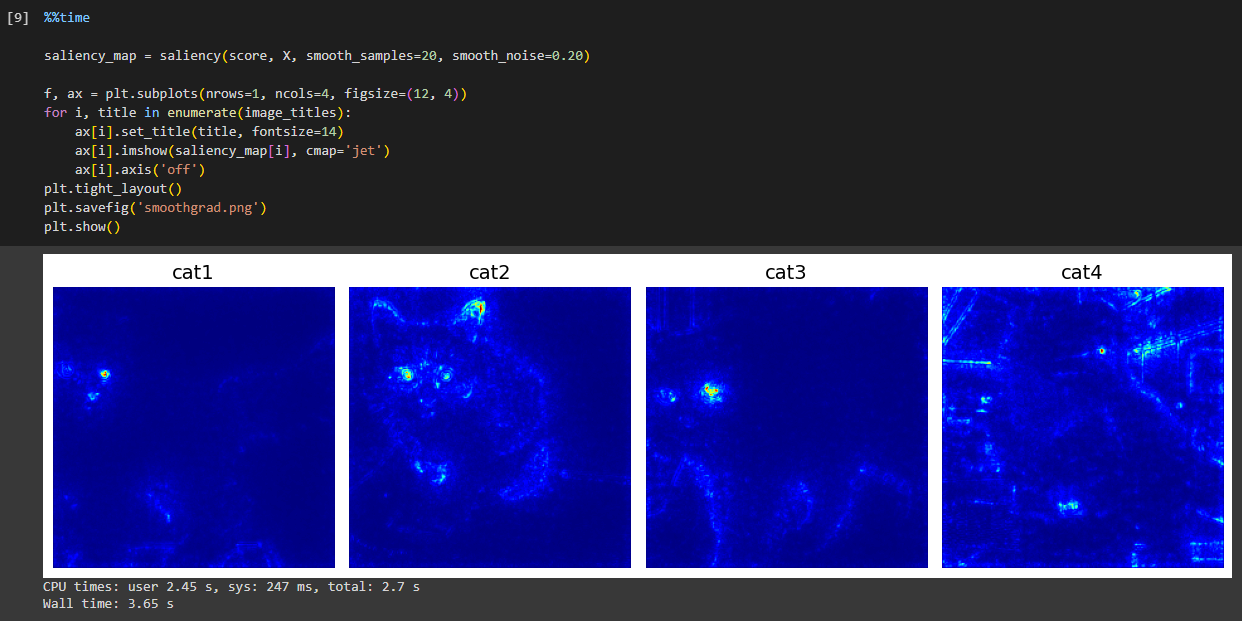


Рисунок – 8

Попробуем способ gradcam. Изначальный обьект виден, но карта явно не охватывает основную цель изображения. Карты показаны на рисунке 9.

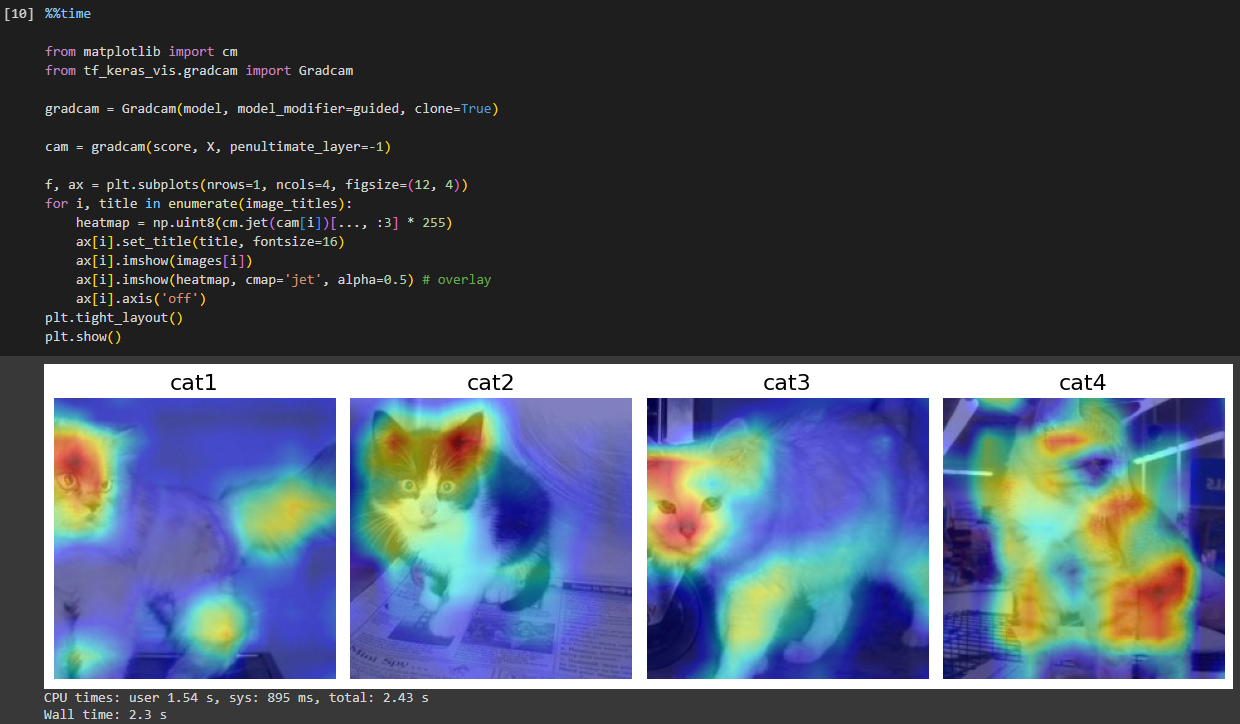


Рисунок – 9

Отобразим gradcam++. Улучшенная версия gradcam практически полностью захватывает объект. Карты показаны на рисунке 10.

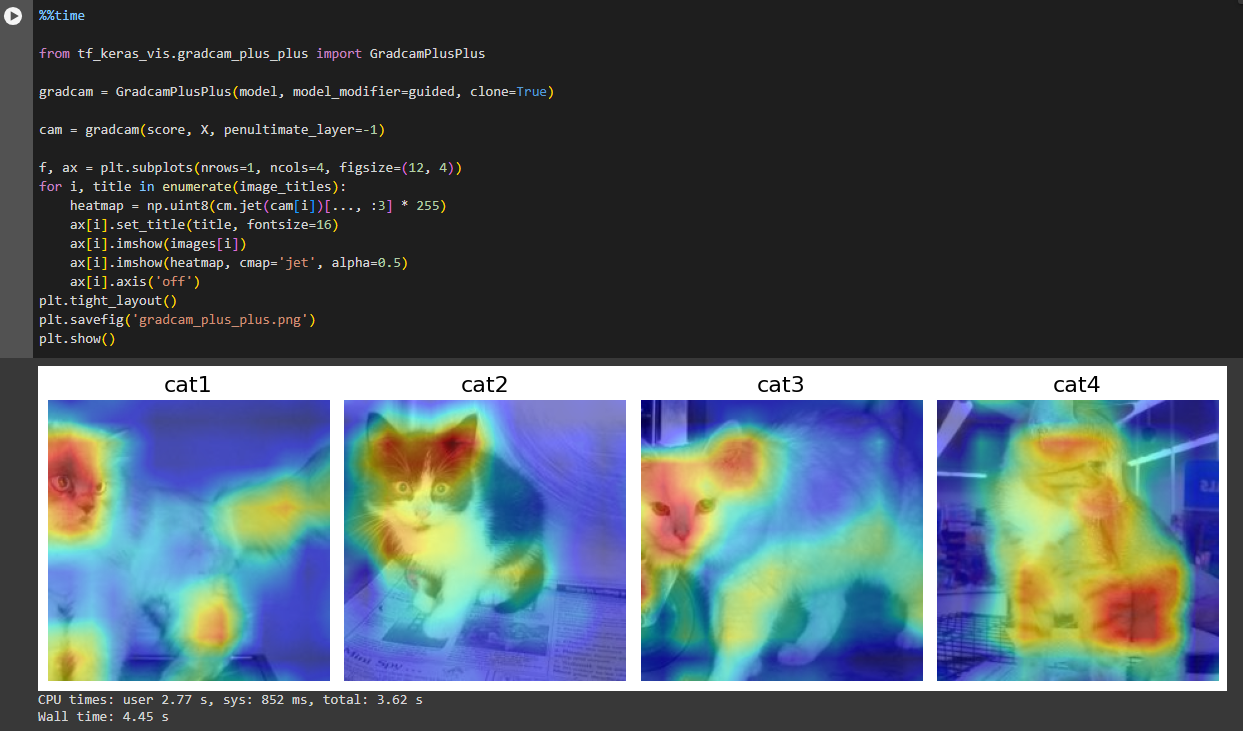


Рисунок – 10

Выводы

В лабораторной работе был разобран процесс построения карт внимания в нейронных сетях для анализа изображений из датасета ImageNet. В ходе работы были выполнены следующие шаги:

Замена функции активации softmax на линейную для корректного вычисления градиентов. Построение карт значимости классов для выбранных изображений методами saliency, smoothgrad, gradcam, gradcam++.

Сравнение результатов и выводы о наиболее точном и полном методе описания активаций слоев нейронной сети.

В результате лабораторной работы были получены информативные карты значимости признаков и классов для изображений из датасета ImageNet.

Это позволило лучше понять, какие части изображений влияют на классификацию, и освоиться с методами построения карт внимания.