

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

**«МИРЭА – Российский технологический университет»** **РТУ МИРЭА**

Кафедра: КБ-4 «Киберразведка и противодействие угрозам с применением технологий искусственного интеллекта»

**Лабораторная работа №3 по дисциплине**

**«Анализ защищенности систем искусственного интеллекта»**

ВыполнилМорин А.А.

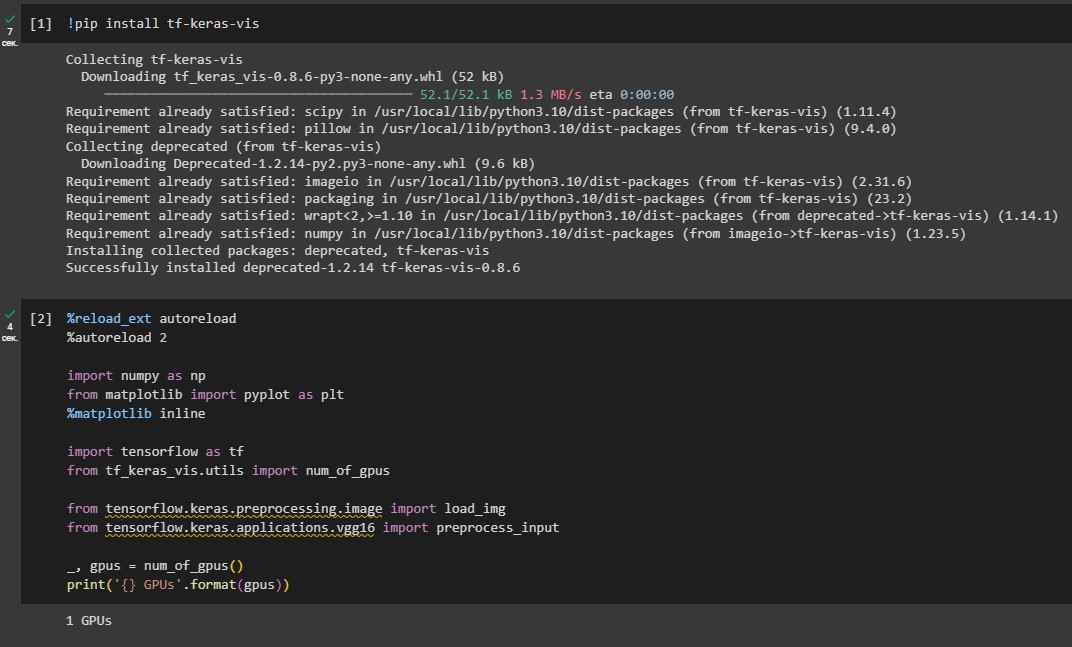
Группа:

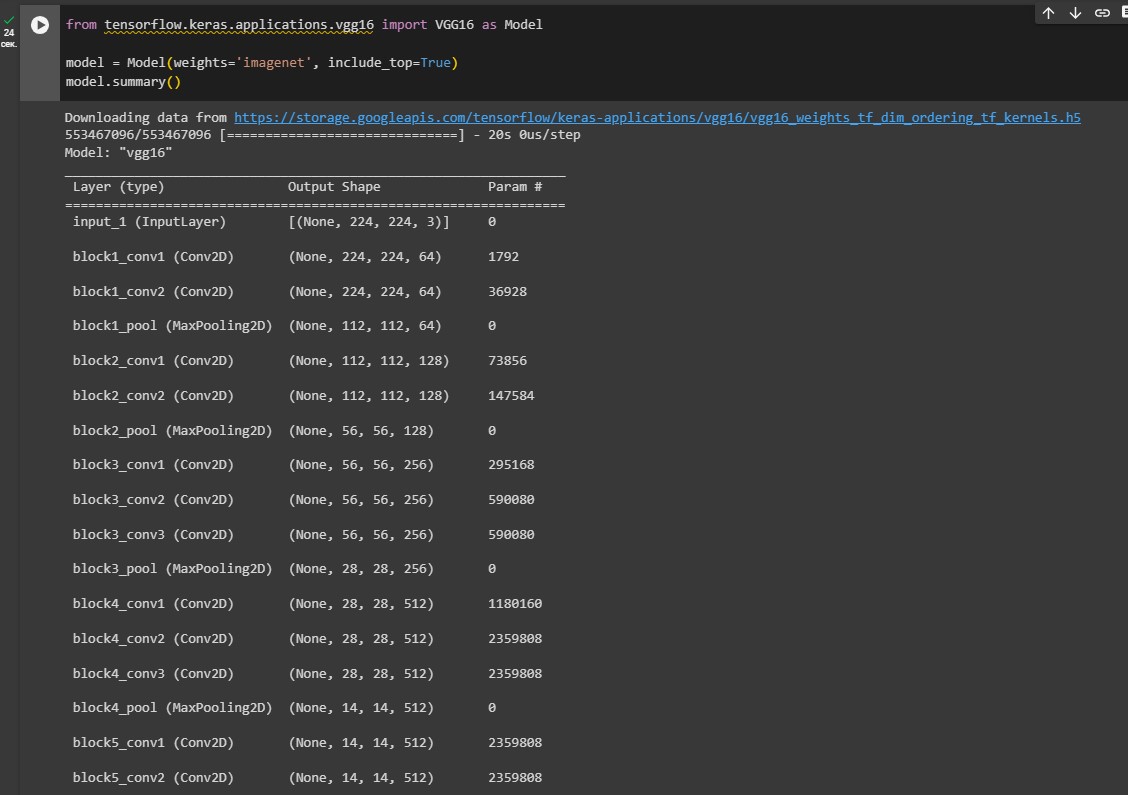
ББМО-01-23

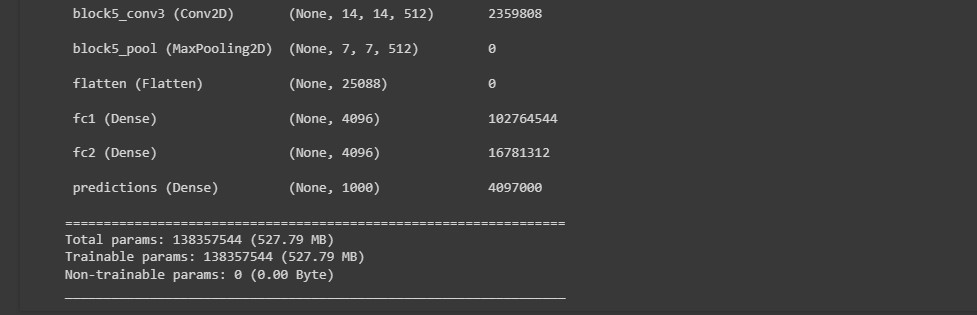
Москва, 2024 г.



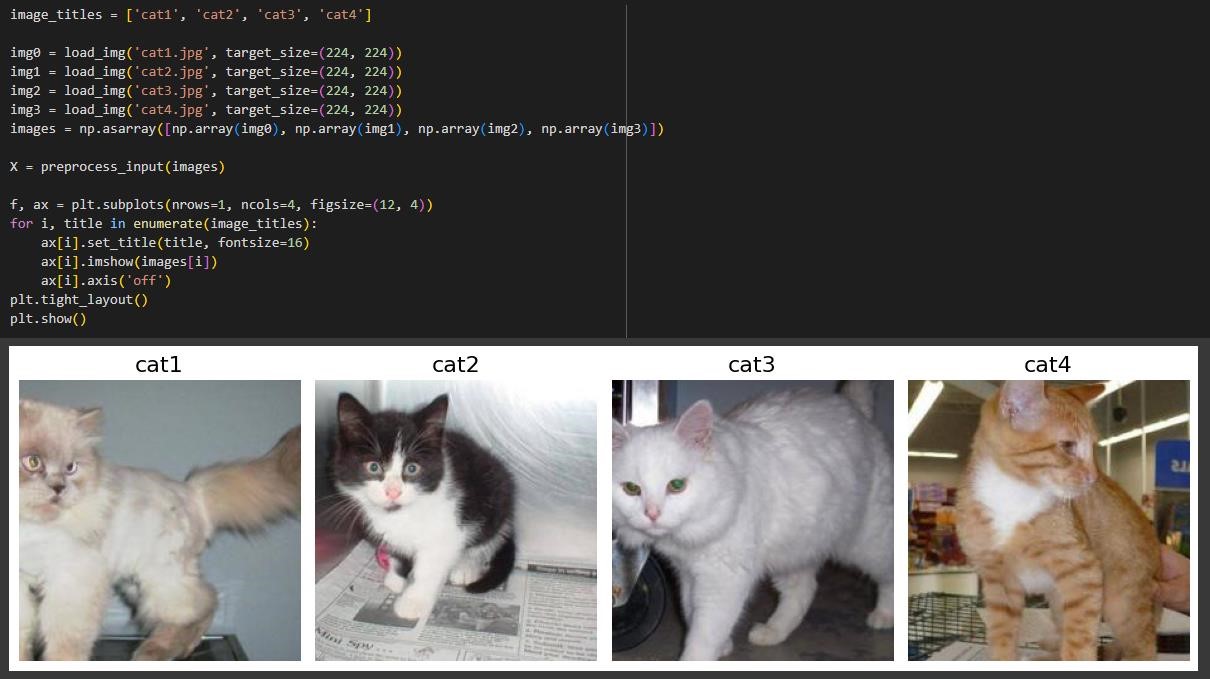
Добавим требующиеся библиотеки и установим keras, a также модель VGG16



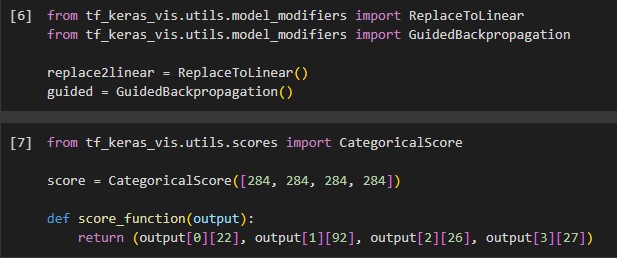




Загрузим и подготовим изображения



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



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



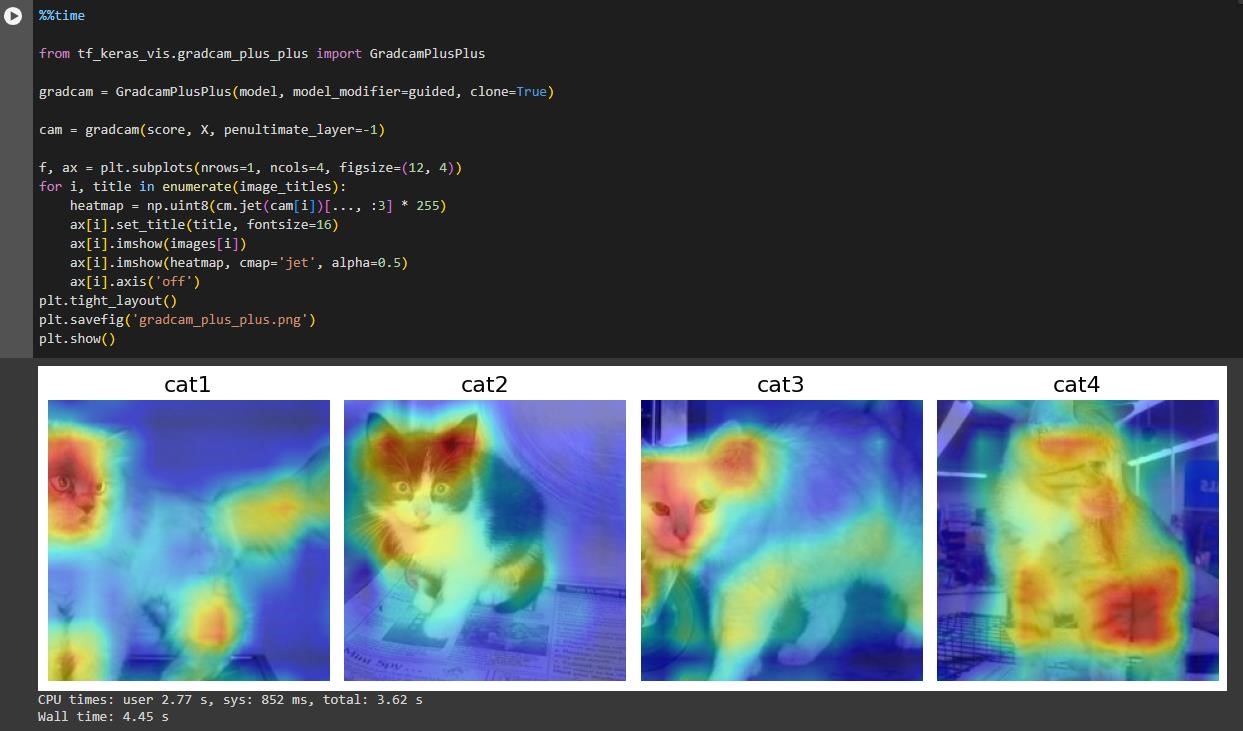
Смотрим карты smoothgrad. Видим улучшенное качество карты, можно понять, что изначальный объект кот.



Попробуем способ gradcam. Изначальный обьект виден, но карта явно не охватывает основную цель изображения.



Отобразим gradcam++. Улучшенная версия gradcam практически полностью захватывает объект.



# **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.