|  |
| --- |
|  |
| 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

Исаев Александр Михайлович

**Проверил**:

К.т.н. Спирин Андрей Андреевич

Москва 2023

# Использование механизмов внимания в нейронных сетях

Для того чтобы успешно выполнить задания, прежде всего необходимо установить tf-keras-vis - реализацией TensorFlow спецификации Keras API.

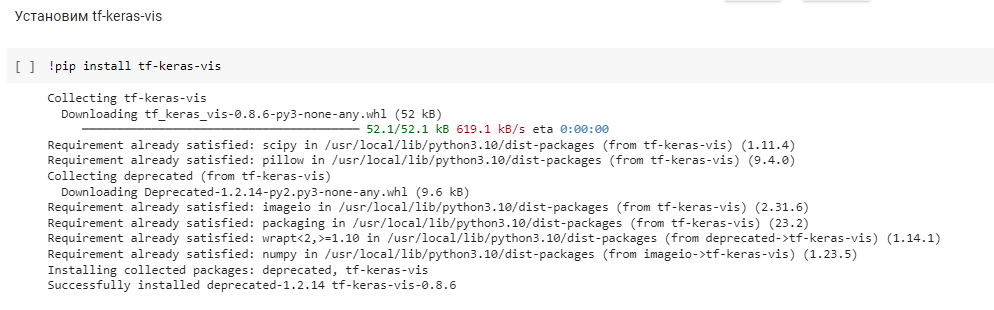


Рисунок 1 – tf-keras-vis

Импортируем библиотеки и распознаем GPU

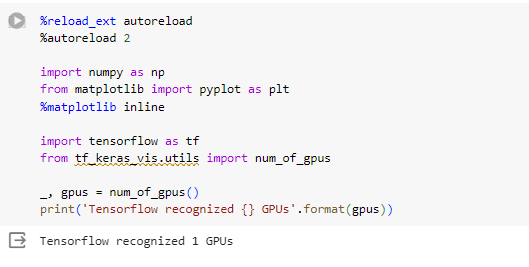


Рисунок 2 – импорт библиотек

Загружаем и импортируем VGG16 – нейросеть для выделения признаков изображений



Рисунок 3 – Используем сеть VGG16

Приступим к 1 части: загрузим кастомные изображения из датасета ImageNet(в другом случае существует вероятность некорректного результата выполенния работы. Стоит отметить что keras-vis позволяет обрабатывать сразу несколько изображений.)и преобразуем их в np.array массивы

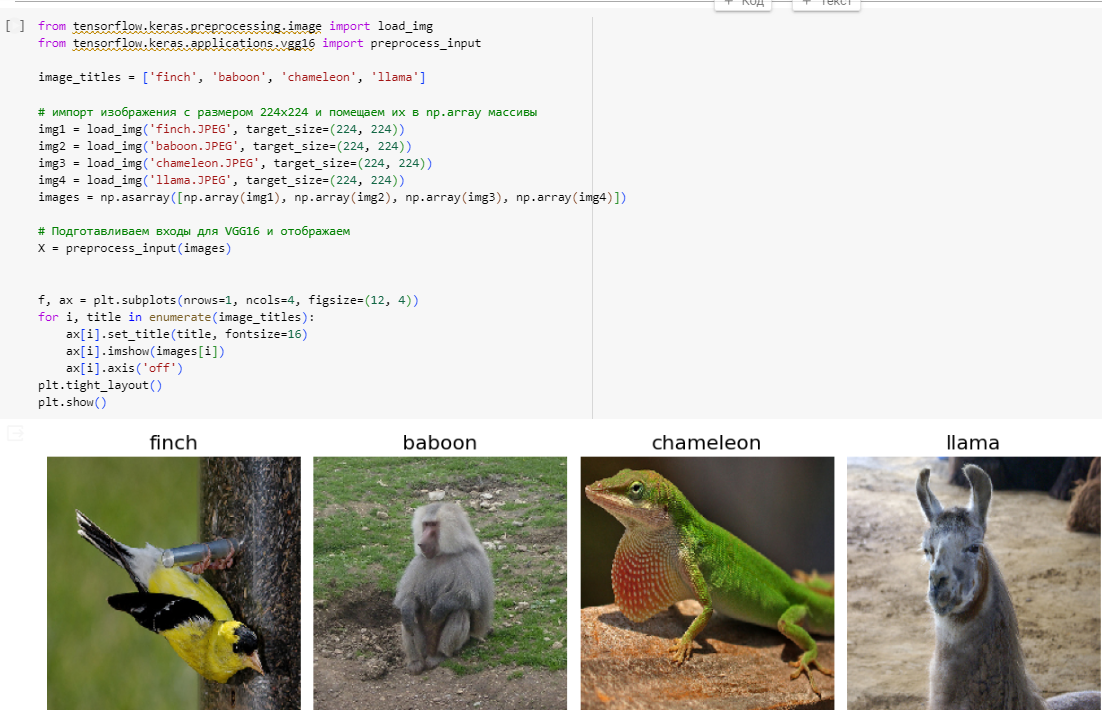


Рисунок 4 – выбираем изображения из датасета

Данные классы имеют 14,372,47,355 индексы, соответственно. Указываем функцию определения значения с индексами.

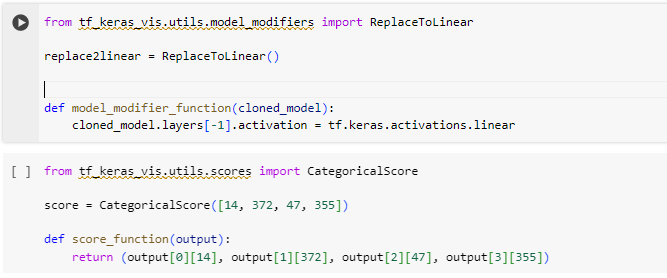


Рисунок 5 – распределяем индексы

Производим генерацию объекта(с помощью функции tf\_keras\_vis.saliency) и карты значимости и отображаем.

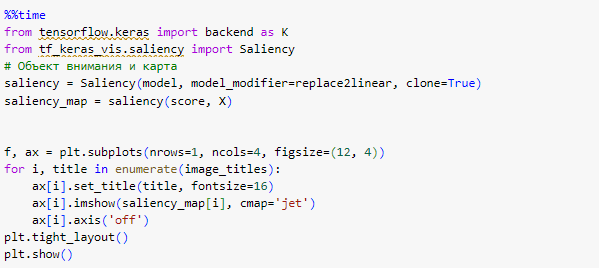


Рисунок 6 – Использование ванильного внимания Saliency

# Saliency

Ванильное внимание Saliency - это метод визуализации, который позволяет понять, какие части входных данных (например, изображений) были наиболее значимыми для принятия решения моделью машинного обучения. Он основан на вычислении градиентов выхода модели по входным данным.

Этот метод позволяет визуализировать, на какие части данных модель "смотрит" при принятии решений, что может помочь в интерпретации работы модели и выявлении ее поведения.

На выходе получаем достаточно сомнительные результаты – имеется общая тепловая карта но с плохим отображением.

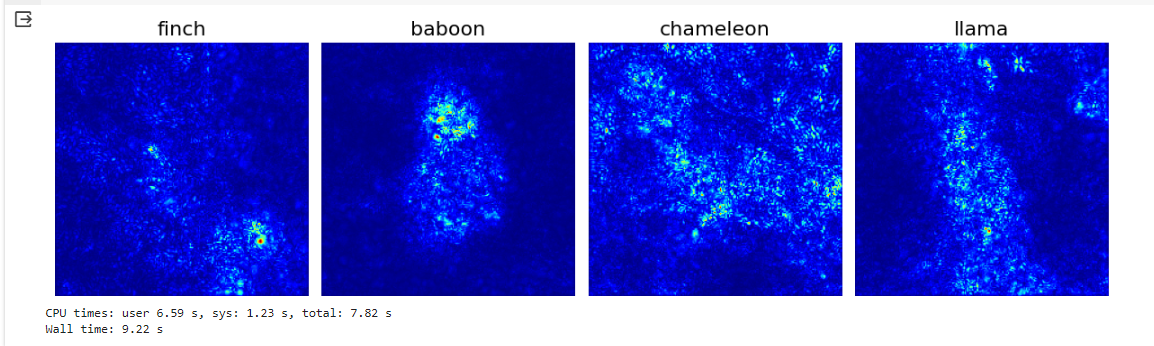


Рисунок 7 – результат выполнения Saliency

# SmoothGrad

SmoothGrad - это метод улучшения визуализации внимания Saliency, который направлен на уменьшение шума и повышение интерпретируемости результатов.

SmoothGrad помогает справиться с проблемой шума в визуализации внимания Saliency, что делает интерпретацию модели более надежной и улучшает понимание того, какие области данных наиболее важны для принятия решений моделью машинного обучения.

Попробуем сменить модификатор на guided и отобразим что получится на выходе. Генерация методом Smoothgrad в итоге отображает более точный итоговый результат.

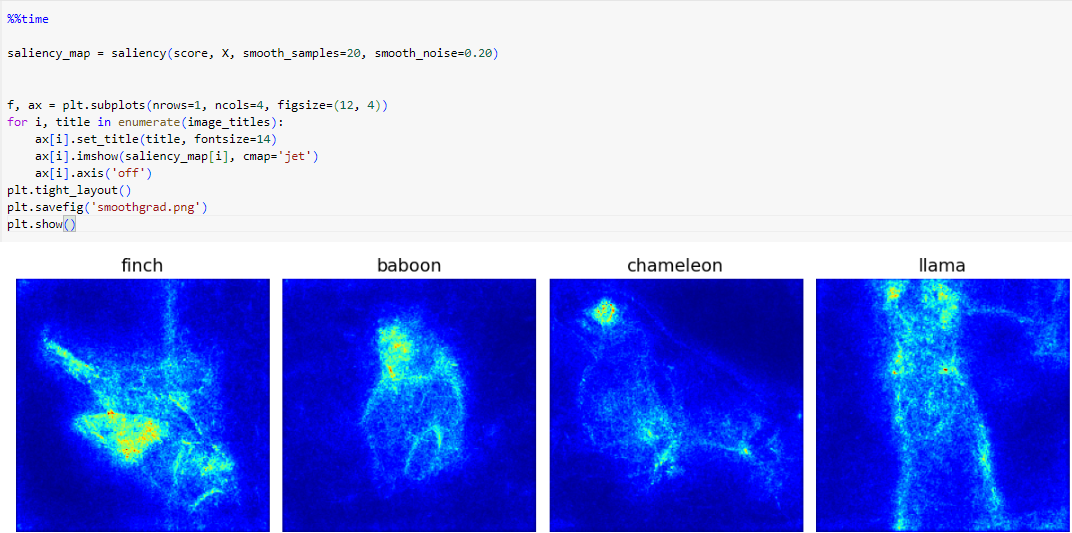


Рисунок 8 – результат выполнения SmoothGrad

# GradCAM

GradCAM (Gradient-weighted Class Activation Mapping) - это метод визуализации, который позволяет понять, какие области изображения были наиболее активными для определенного класса при принятии решения моделью глубокого обучения. Он расширяет идею внимания Saliency, фокусируясь не только на вычислении градиентов выхода модели по входным данным, но и на весовых коэффициентах, которые объединяют градиенты с активациями слоев модели. GradCAM позволяет визуализировать активации, ассоциированные с различными классами объектов на изображениях.

Использует модуль tf\_keras\_vis.gradcam для создания GradCAM (Gradient-weighted Class Activation Mapping) для модели машинного обученияЭтот метод создает тепловые карты, показывающие, какие области входных данных были наиболее активными для определенных классов.

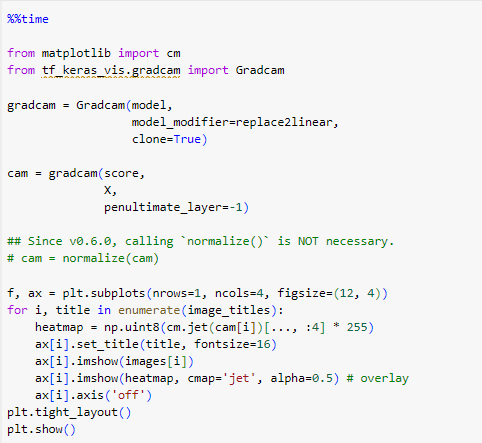


Рисунок 9 - GradCAM

Результат выполнения показан на следующем рисунке.

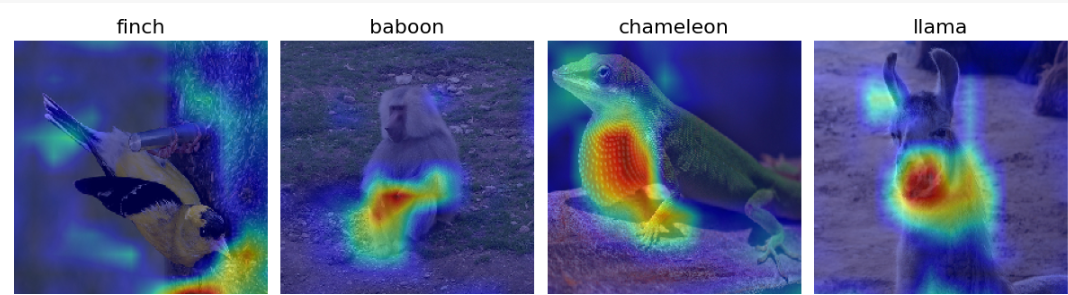


Рисунок 10 – результат выполнения GradCAM

# GradCAM++

GradCAM++ (Gradient-weighted Class Activation Mapping ++) является улучшенной версией GradCAM, направленной на более точное определение активаций объектов на изображениях.

Основное улучшение GradCAM++ состоит в учете не только положительных градиентов (которые указывают на активацию объектов), но и отрицательных градиентов (которые указывают на отсутствие объектов) для лучшего выделения объектов на изображениях.

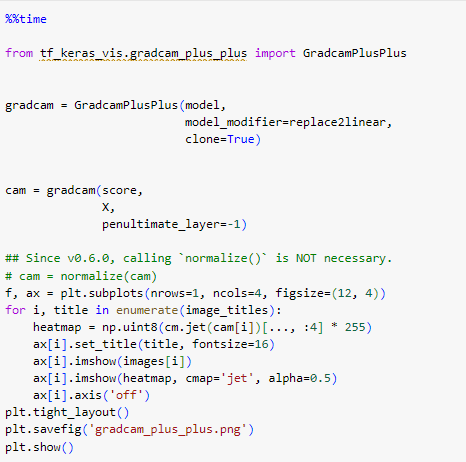


Рисунок 11 – GradCAM++

Результат выполнения показан на следующем рисунке.

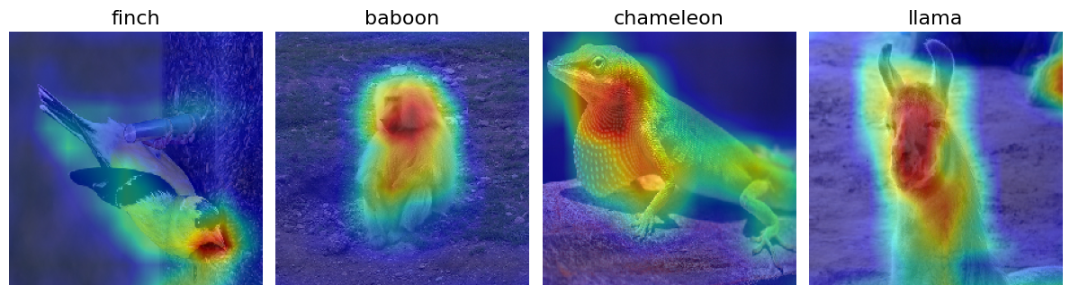


Рисунок 12 – результат выполнения

# **ВЫВОД**

Таким образом, в результате проведения лабораторной работы 3 были выявлены навыки работы с различными инструментами получения карт активации, методы создания карт активации – и их модификаций.