Разработка алгоритма контрфактических объяснений в мультимодальных моделях BLIP и CLIP

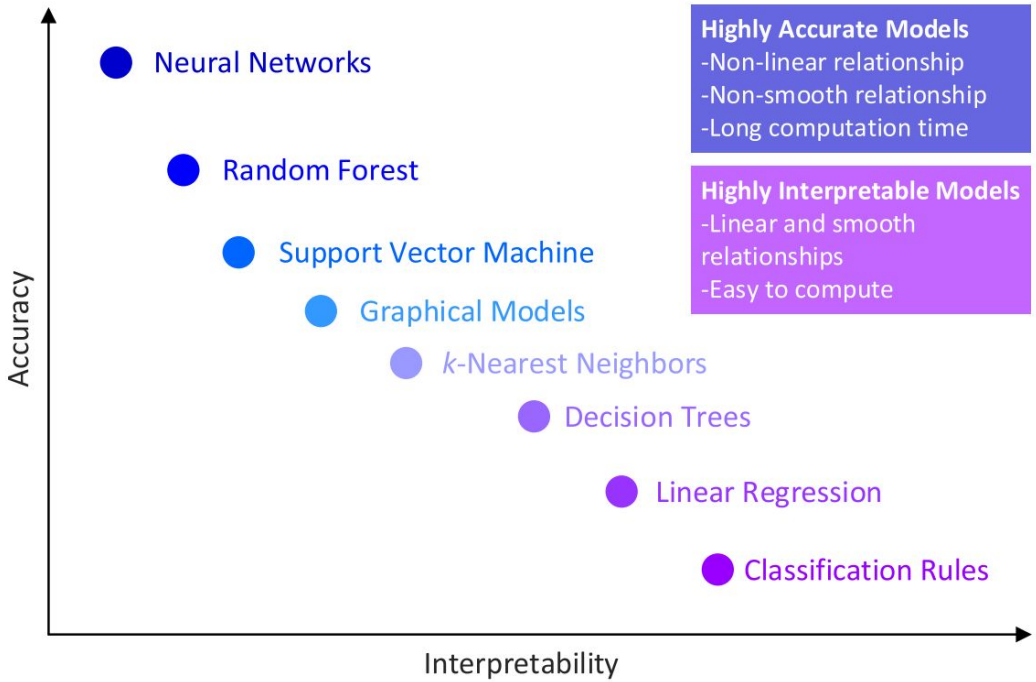
Студент: Кондренко Кирилл Павлович
Руководитель ВКР: Ряскин Александр Николаевич

Новосибирск 2025

Актуальность

- Широкое применение мультимодальных моделей
- Сложность объяснения решений глубоких моделей
- Во многих областях от моделей требуется объяснимость
- Объяснимость и точность моделей коррелируют отрицательно

Объяснимость и точность моделей



Цели и задачи

Цель: разработать алгоритм генерации контрфактических объяснений для мультимодальных моделей BLIP и CLIP.

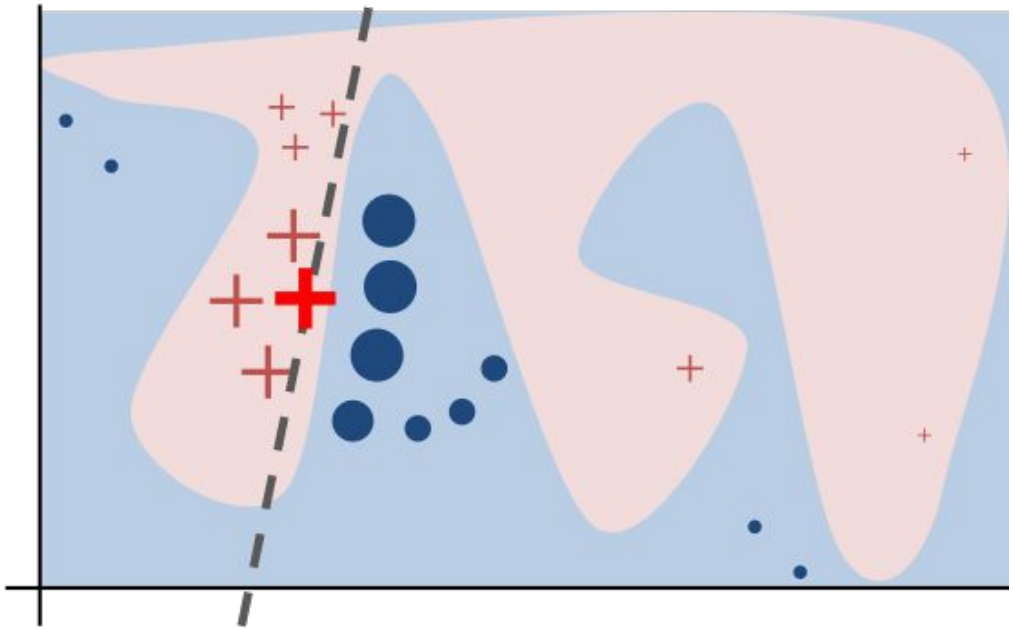
Задачи:

1. Проанализировать решения в области объяснимого ИИ.
2. Разработать алгоритм генерации контрфактических объяснений для моделей BLIP и CLIP.
3. Провести оценки объяснений, генерируемых алгоритмом.

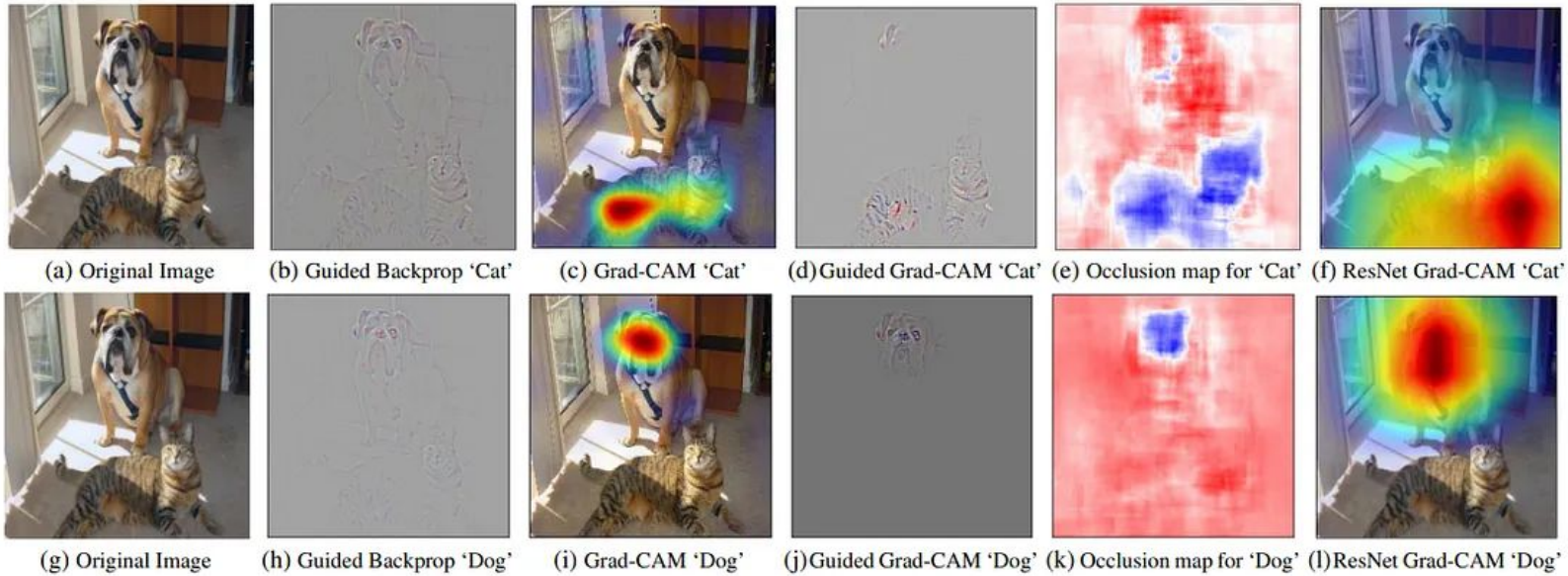
Подходы к объяснениям

- Через модель
- Аппроксимация модели в окрестности точки данных с помощью более простой (LIME)
- Оценка вклада признаков в предсказание (SHAP, Grad-CAM)
- Контрфактические объяснения — минимальные изменения, приводящие к другому предсказанию модели
- Гибридные

LIME



Grad-CAM



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

CLIP

- Модель работает с текстом и изображениями в одном векторном пространстве
- Для пары (текст, изображение) выдаёт их семантическую схожесть
- Демонстрирует высокие показатели на изображениях и текстах, которых не было при обучении

Применение CLIP

Классификация:

- Подаём на вход изображение и несколько текстов (классов)
- Модель относит изображение к одному из данных классов с некоторой степенью уверенности

Фильтрация:

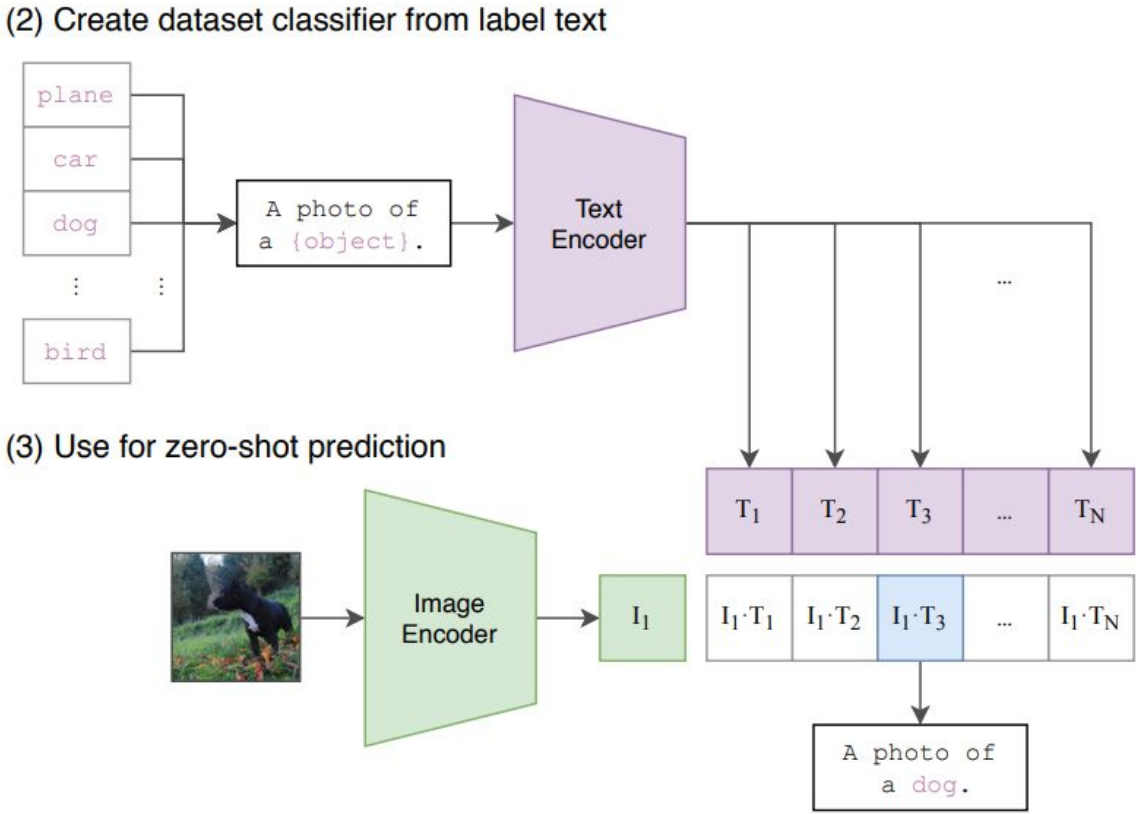
- Подаём на вход изображение и несколько текстов (классов)
- Для каждого класса модель возвращает степень уверенности того, насколько класс есть в изображении
- Если степень уверенности > порога, то изображение отфильтровывается

Применение CLIP

Поиск:

- Подаём на вход текст
- Модель генерирует эмбеддинг текста
- Ищем в векторной базе данных изображения, эмбеддинги которых близки к эмбеддингу текста

Пример классификации



Объяснения для CLIP

- Нужно научиться для пары (текст, изображение) отвечать на вопрос, почему именно такая степень уверенности?
- **От изображения**: как изображение повлияло на степень уверенности?
- **От текста**: как текст повлиял на степень уверенности?

От изображения

Контрфактические объяснения:

1. С помощью обратного хода определяем, какой должен быть эмбеддинг e изображения, чтобы увеличить степень уверенности.
2. Ищем в векторной базе данных изображения, эмбеддинги которых близки к e.
3. Если для некоторого изображения степень уверенности стала нужной, то возвращаем его.

Важность признаков: С помощью обратного хода определяем и подсвечиваем пиксели изображения, которые больше всего повлияли на выход модели (нужно адаптировать существующие подходы под визуальный трансформер).

От текста

Контрфактические объяснения:

1. С помощью обратного хода определяем, какой должен быть эмбеддинг *e* текста, чтобы увеличить степень уверенности.
2. Ищем в векторной базе данных тексты, эмбеддинги которых близки к *e*.
3. Если для некоторого текста степень уверенности стала нужной, то возвращаем его.

Важность признаков: С помощью обратного хода определяем и подсвечиваем слова текста, которые больше всего повлияли на выход модели (нужно адаптировать существующие подходы под трансформер).

Качество объяснений

- **Верность** — насколько объяснение отражает реальное поведение модели (изменить изображение/слова класса и посмотреть изменение результата модели)
- **Понятность** — насколько объяснение легко понять человеку
- **Полнота** — Отражает ли объяснение все существенные факторы
- **Стабильность** — объяснения должны быть непрерывной функцией от входных данных
- **Контрастность** – почему выбран именно этот класс и что сделать, чтобы результат стал другим

Заключение

1. Проанализированы решения в области объяснимого ИИ.
2. Спроектирован алгоритм генерации ~~контрфактических~~ объяснений для модели CLIP.