Разработка мультимодальных нейросетевых методов распознавания изображений на основе трансформерных архитектур

Авшалумов М.С.
avshalumov.ms@phystech.edu

Центр когнитивного моделирования МФТИ

18 марта 2025

Мотивация

Традиционные методы распознавания изображений часто сталкиваются с ограничениями:

- При работе со сложными сценами, например с большим количеством объектов, или содержащими реальные данные, где объекты могут иметь схожие визуальные характеристики или быть частично видимыми
- При работе с объектами, которые не наблюдались в процессе обучения
- В задачах реального применения затруднительно подбирать свой визуальный запрос(точку, bounding box, маску) для каждого целевого изображения

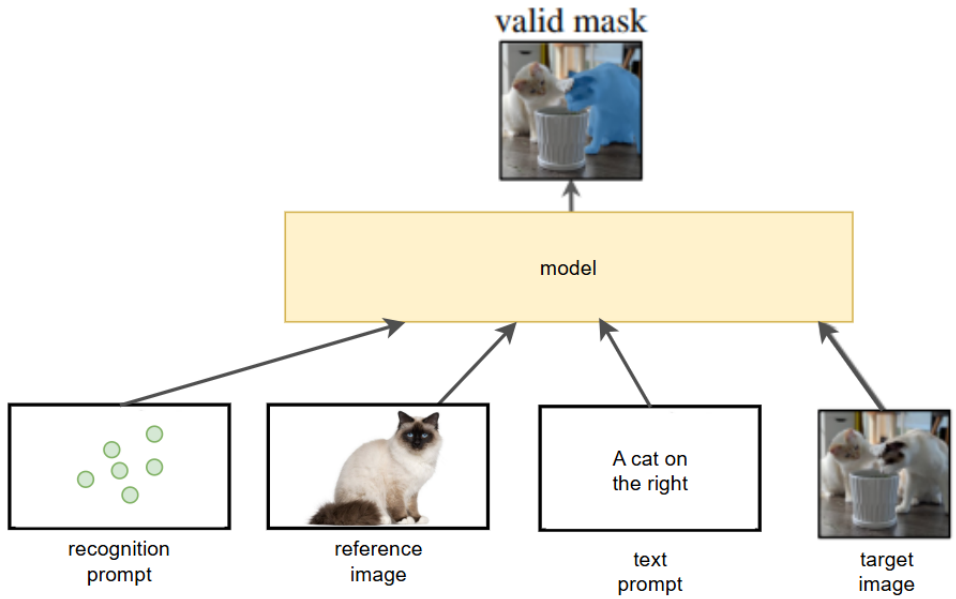




Примеры сложных для распознавания сцен

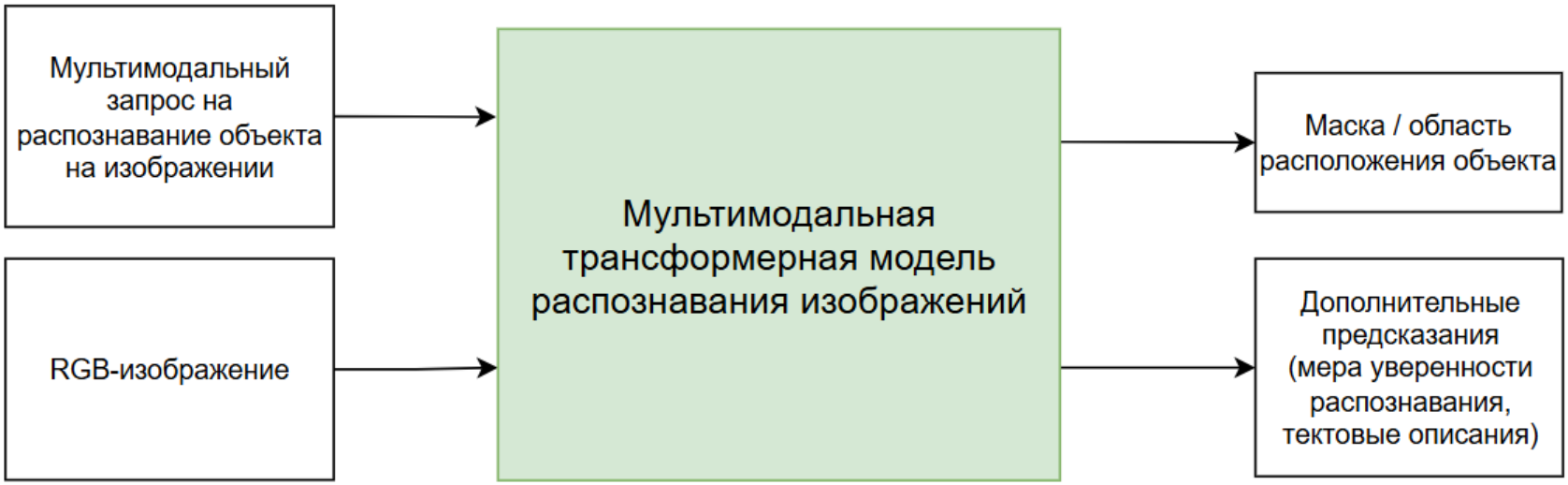
Мотивация

Для улучшения качества работы применяется различная мультимодальная информация, такая как дополнительные изображения, маски или текстовые описания, что позволяет учитывать дополнительный контекст



Цель работы

- Целью данной работы является разработка эффективных и высокопроизводительных нейросетевых методов распознавания изображений по мультимодальным запросам с использованием архитектур типа трансформер



Предполагаемая схема работы

Постановка задачи

- **Входные данные**:
  - I: входное изображение, представленное как тензор размера $H \times W \times C$, где H — высота, W — ширина, C — количество каналов (например, 3 для RGB).
  - Q: тензор размера $H_q \times W_q \times C_q$, представляющий собой мультимодальный запрос на распознавание.
- **Целевая переменная**:
  - M: маска сегментации, представляющая собой бинарную матрицу размера $H \times W$, где $M_{ij} = 1$, если пиксель (i, j) принадлежит целевому объекту, и $M_{ij} = 0$ в противном случае.
  - A: дополнительные выходы модели, такие как текстовые описания, мера уверенности распознавания, классификация объектов на изображении и тд.

Формально, задача состоит в нахождении отображения:

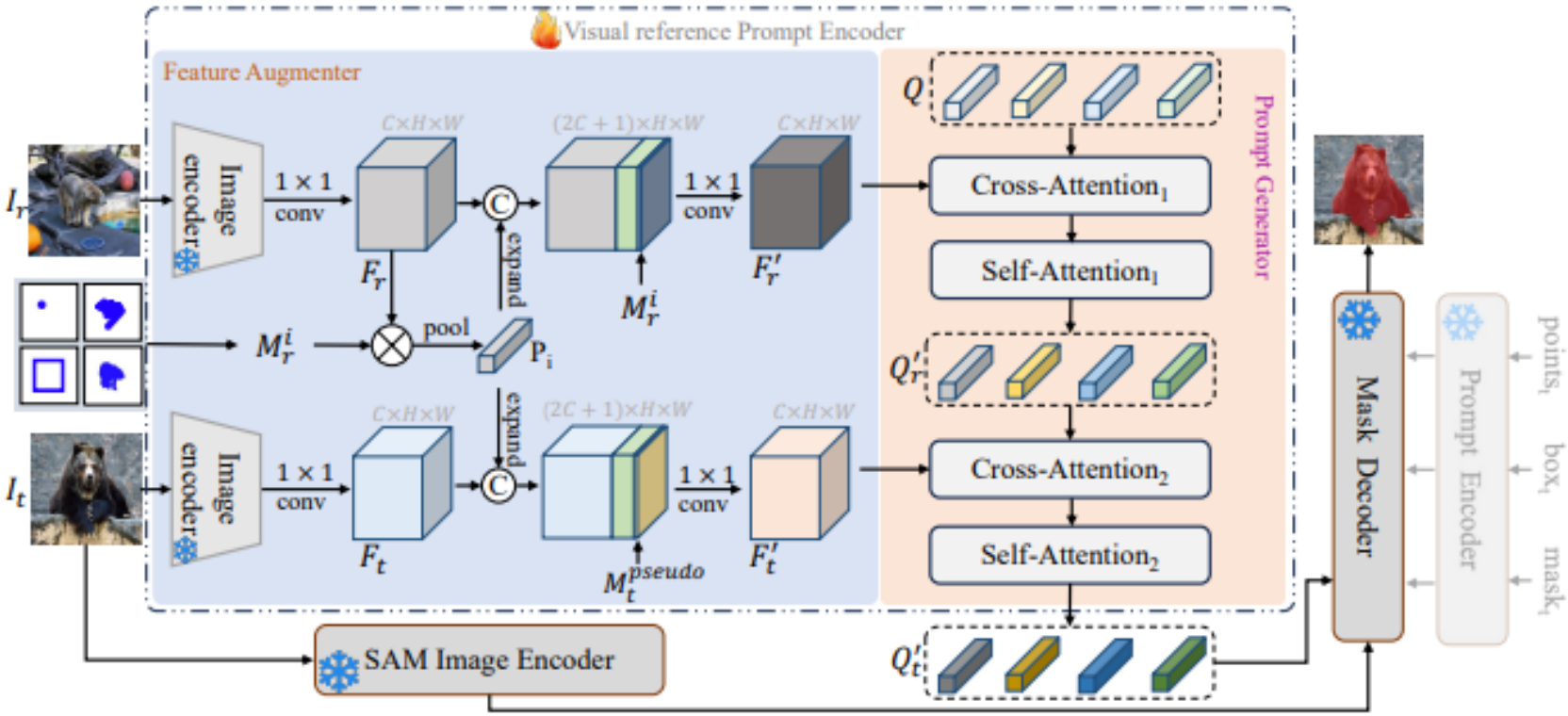
$$f : (I, Q) \underset{L \to 0}{\mapsto} M, A$$

где f — нейронная сеть, а L - функция потерь.

Бейзлайны

Модель VRP-SAM

В качестве бейзлайна для работы с мультимодальными запросами выбрана модель VRP-SAM[1]:



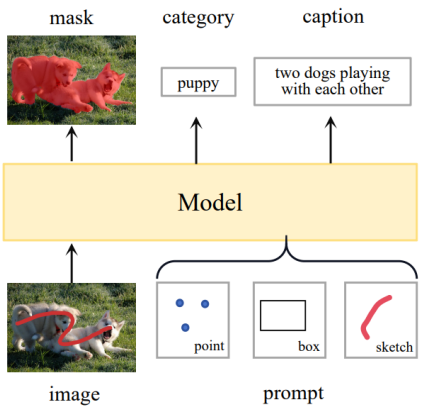
Архитектура модели VRP-SAM

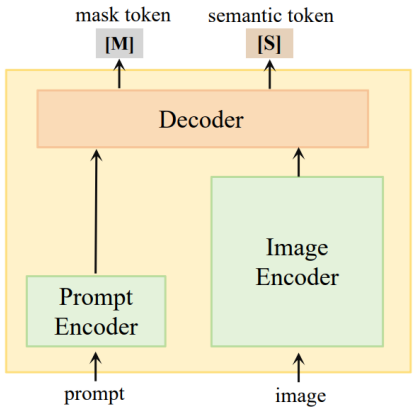
[1] Sun Y. et al. VRP-SAM: SAM with visual reference prompt //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – С. 23565-23574.

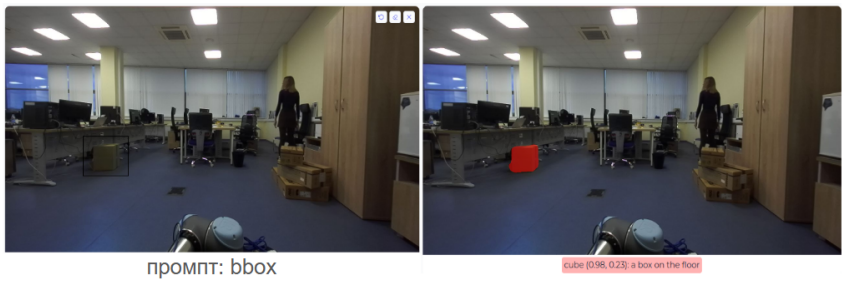
Бейзлайны

Tokenize Anything

В качестве бейзлайна для предсказания масок выбрана модель Tokenize Anything(TAP)[1]:







промпт: bbox



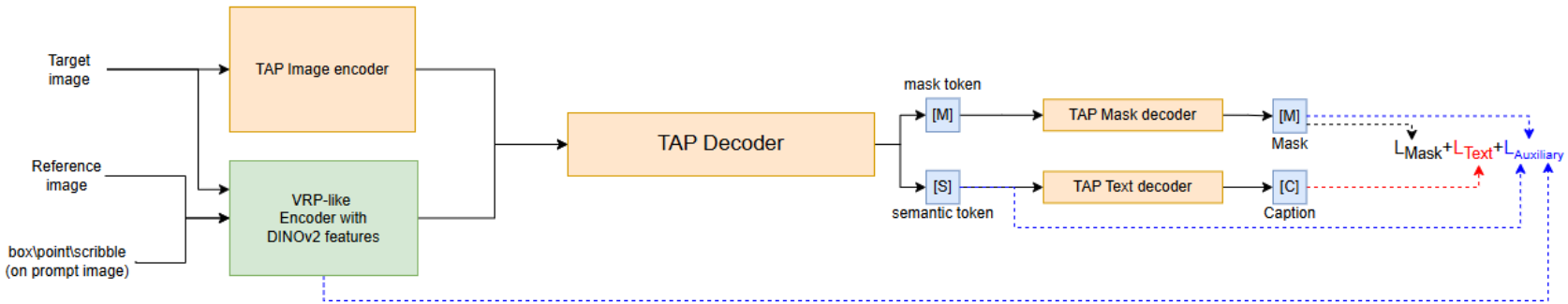
промпт: точки

[1] Pan T. et al. Tokenize anything via prompting //European Conference on Computer Vision. – Cham : Springer Nature Switzerland, 2024. – С. 330–348.

Предлагаемый подход

Ключевые нововведения:

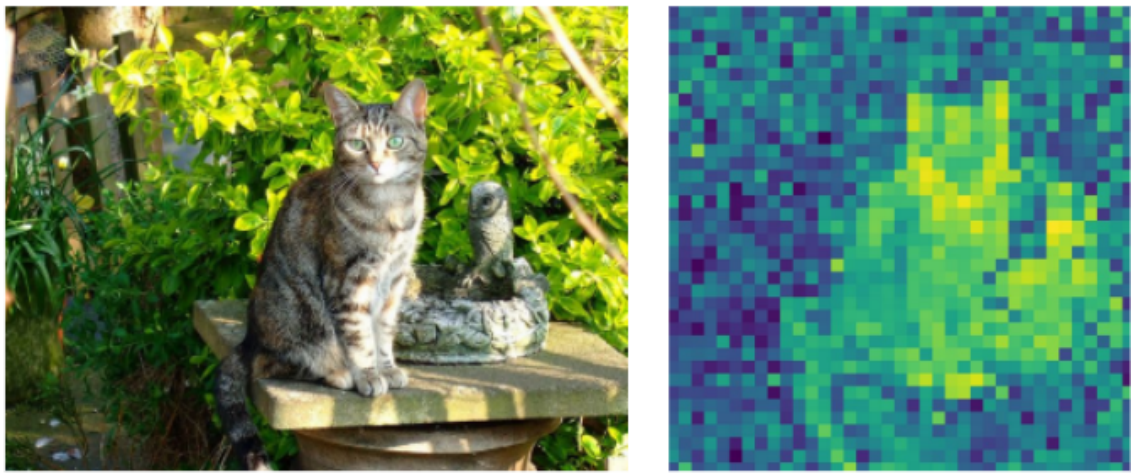
- Новая постановка задачи
- Новый метод обучения (онлайн-выбор референсных изображений и промптов к ним во время обучения)
- Новая функция потерь, с учетом как визуальной, так и текстовой составляющей вывода модели
- Использование DINO признаков в функции потерь для снижения шума масок сегментации (одной из ключевых проблем метода VRP-SAM)



Предлагаемый подход

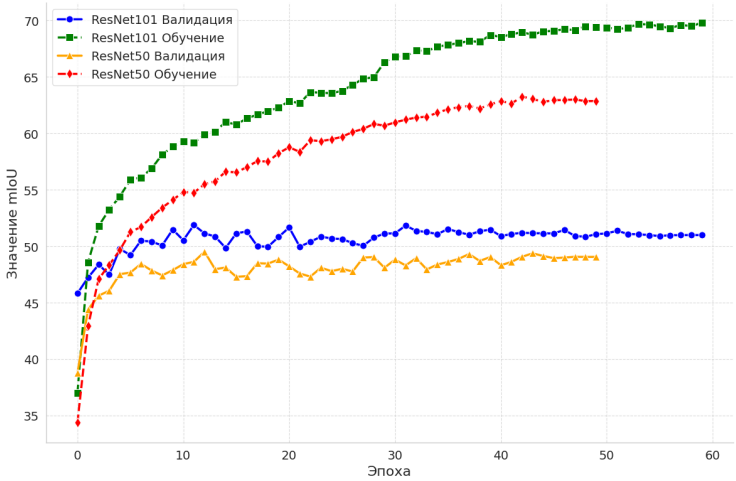
Дополнительная функция потерь:

- Особенность пространства DINOv2 в разделенности признаков разных классов в этом пространстве
- Это позволяет использовать эти признаки для локализации объекта конкретного класса на изображении

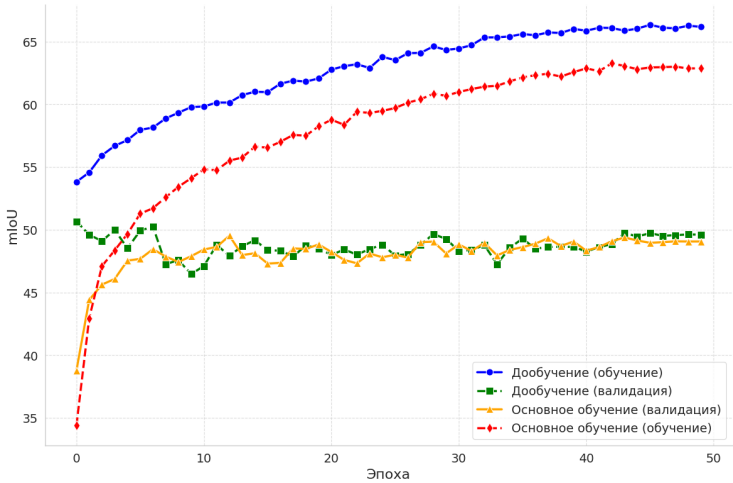


Пример Dinov2 Similarity Map

Эксперименты



(a) Сравнение значений mIoU для моделей VRP-SAM с разными базовыми моделями извлечения признаков



Сравнение значений mIoU для моделей VRP-SAM с дообучением и без

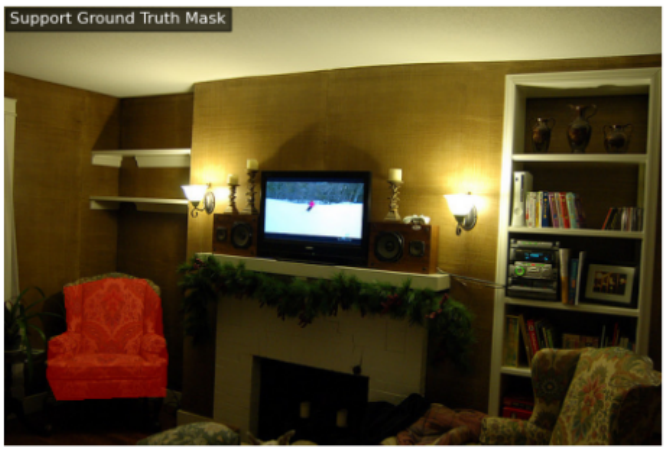
Графики обучения модели VRP-SAM

Эксперименты

Качественные примеры визуализаций работы предлагаемого алгоритма:









Эксперименты

Быстродействие

Результаты замеров быстродействия предложенной архитектуры в сравнении с оригинальной моделью TAP:

Модель	**Тип запроса**	**Быстродействие**
Оригинальная модель TAP	Целевое изображение и промпт на нём	1.3с
Предложенный алгоритм	Целевое и референсное изображения и визуальный промпт на референсном	1.7с

Замеры проводились на видеокарте Nvidia A100 для изображений размером 512x512

Дальнейшие планы

- Качественный анализ работы предложенного алгоритма
- Исследование влияния отдельных компонентов функции потерь на результаты модели
- Исследование использования DINOv2 в качестве модели извлечения признаков для энкодера визуальных запросов

Спасибо за внимание!

avshalumov.ms@phystech.edu