Разработка мультимодальных нейросетевых методов распознавания изображений на основе трансформерных архитектур

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

Центр когнитивного моделирования МФТИ

18 ноября 2024

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

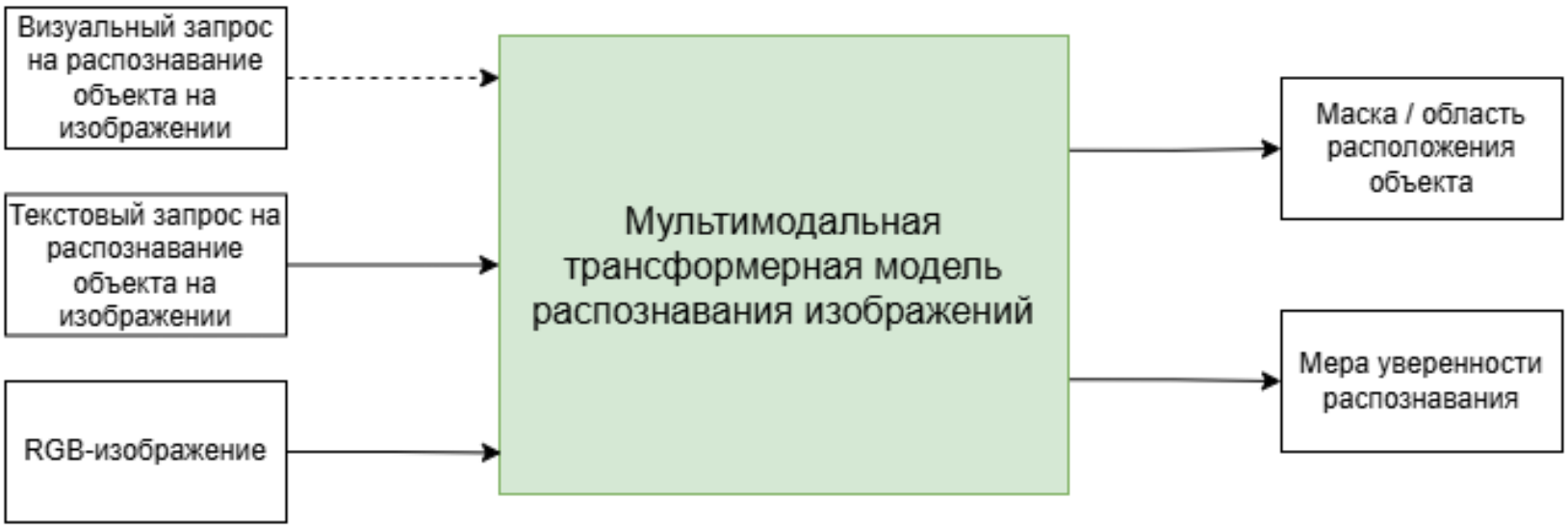




Примеры сложных для распознавания сцен

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

Разработка метода повышения качества распознавания изображений по текстовому запросу

Модели распознавания изображений с открытым словарем

- Это модели машинного обучения, используемые для обнаружения объектов или классификации данных в условиях, когда наблюдаемые объекты могут не принадлежать известным заранее классам.
- Примерами моделей детекции с открытым словарем являются Grounding DINO[1] и YOLO-World[2]
- Примерами моделей сегментации с открытым словарем являются ReferFormer[3] и UNINEXT[4]

---

[1] S. Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection

[2] Cheng, Tianheng et al. : YOLO-World: Real-Time Open-Vocabulary Object Detection

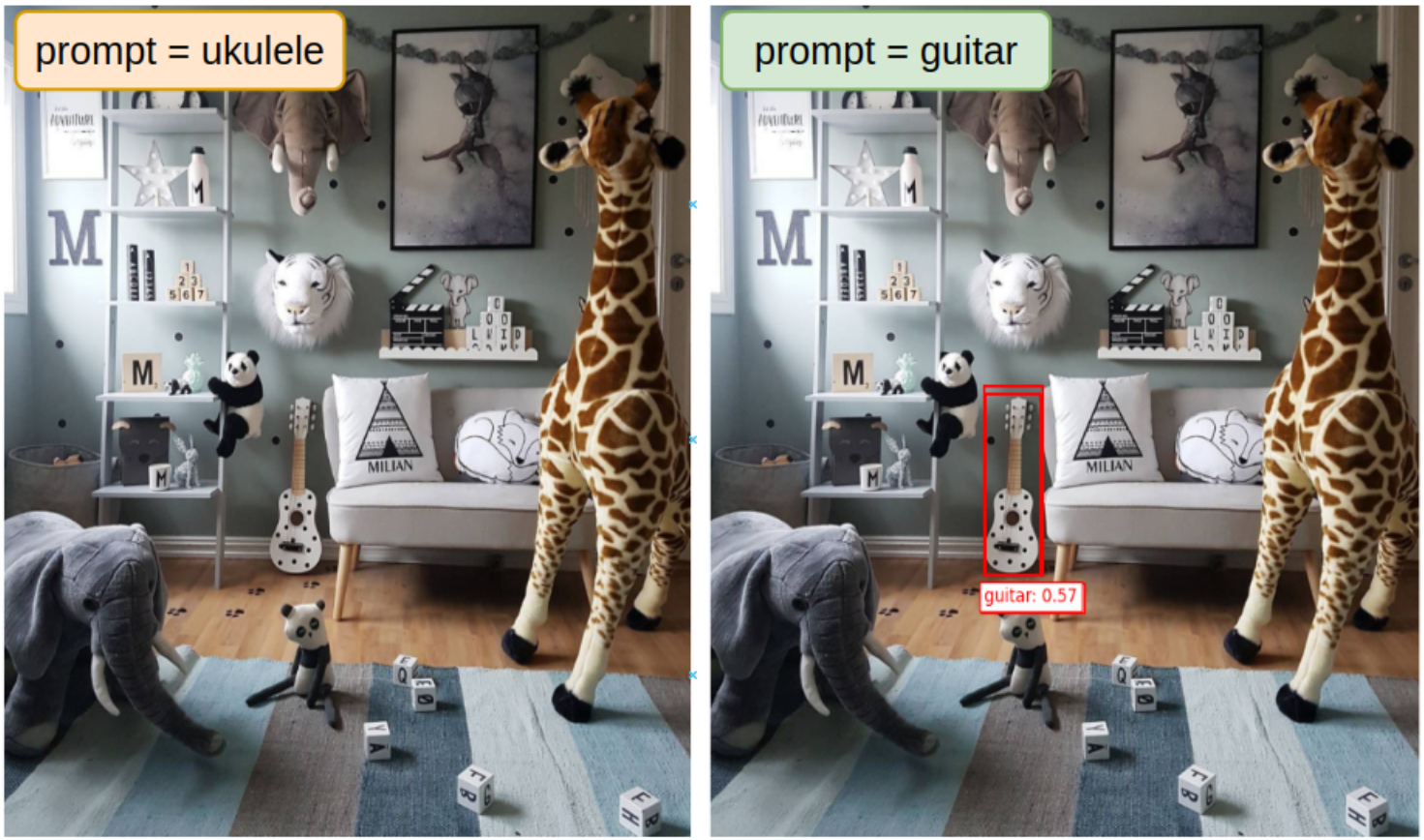
[3] Jiannan Wu et al.: Language as Queries for Referring Video Object Segmentation(2022)

[4] Yan Bin et al.: Universal Instance Perception as Object Discovery and Retrieval(2023)

Мотивация

Формулировка проблемы

- Модели распознавания изображений могут показывать сильно разнящиеся результаты на паре синонимичных с точки зрения натурального языка запросов



Пример синонимичных запросов с разным качеством в задаче детекции

Мотивация

Существующие методы решения

На сегодняшний день существует три возможных способа решить данную проблему без участия человека:

- Автоматический промпт-инжениринг(prompt engineering)
- Дообучение языковой модели для соответствия предпочтениям
- Дообучение мультимодальной трансформерной модели

Мотивация

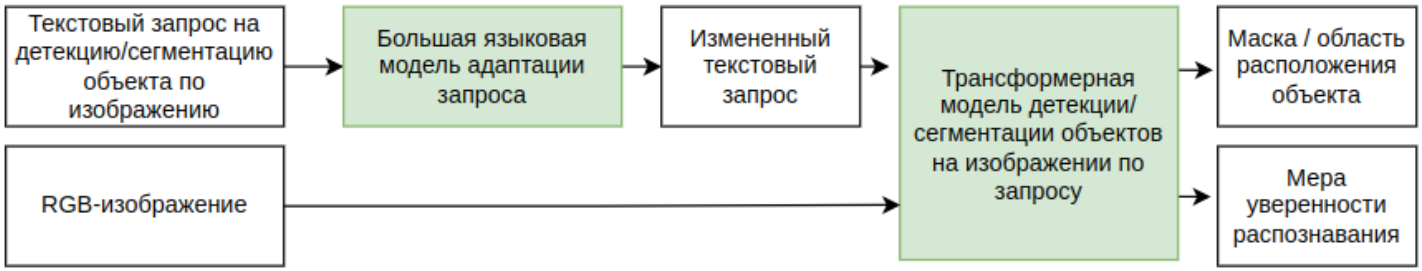
Обсуждение

Существующие подходы сталкиваются со следующими проблемами:

- Используют косвенные оценки качества подобранного промпта
- Требуют шаблона промпта для подбора
- Вычислительно сложны и требуют большого количества данных

Дообучение языковой модели для оптимизации промптов с использованием отклика от модели детекции/сегментации позволит решить эти проблемы

Предлагаемый метод



Предполагаемая схема работы

Постановка задачи

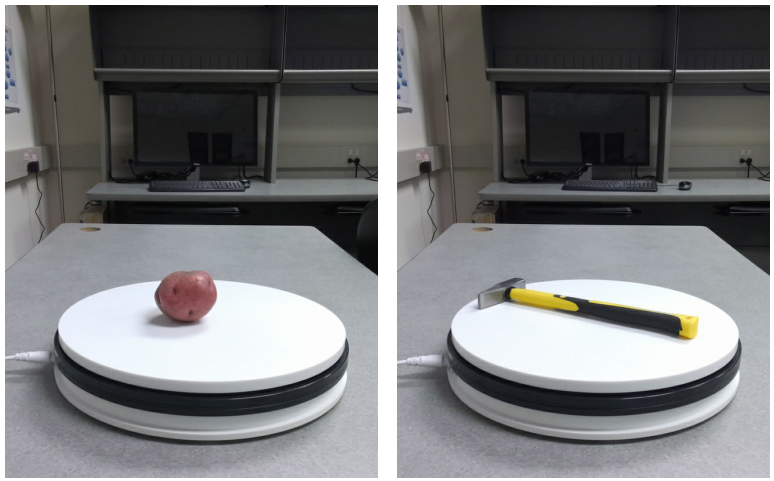
Дообучение языковых моделей как MDP

Сформулируем MDP для решаемой задачи:

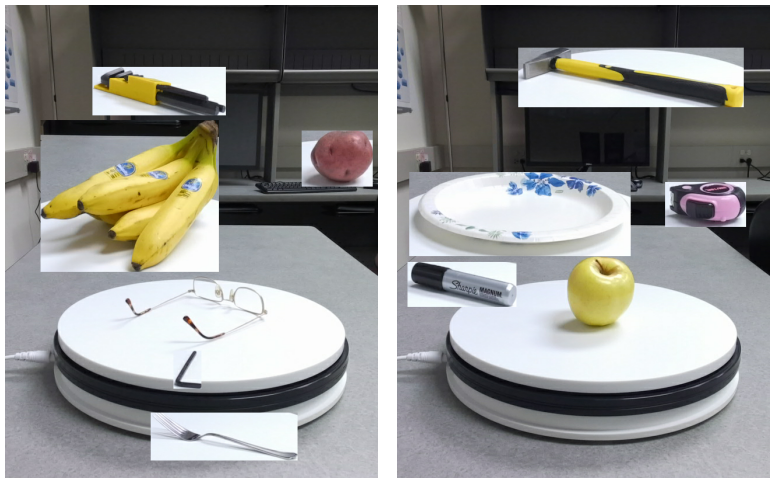
- *Пространство состояний* S: область всех возможных предсказанных токенов, состояние $s \in S$ - текущая текстовая последовательность;
- *Пространство действий* A: ответы модели на запросы, действие $a \in A$ - предсказание следующего токена;
- *Функция вознаграждения* R: некоторая функция от выводов модели детекции;
- *Стратегия* $\pi(s, \theta)$: сама языковая модель, которая каждому состоянию(текущей последовательности токенов) сопоставляет вероятностное распределение действий(возможных следующих токенов);

Датасет

С целью увеличения разнообразия входных изображений была произведена аугментация базового набора данных GOLD[1]. Полученный набора данных был назван GOLD-A(Augmented)



(a) Изображения из датасета GOLD



(b) Изображения из датасета GOLD-A

Примеры изображений из датасетов GOLD и GOLD-A

[1]Kebe, G.Y. et al.: A spoken language dataset of descriptions for speech-based grounded language learning (2021)

Датасет

Датасет содерит 825 изображений размером 640x810 пикселей и более 300 тысяч текстовых описаний

Структура датасета

Название	**Описание**	**Пример**
item id	Название класса, к которому принадлежит целевой объект на изображении. Однозначно определяет, к какому изображению относится описание	*fork_1_2*
true bbox	Истинный bbox для целевого объекта	[236, 286, 231, 328]
description	Описание объекта	*A fork designed for picking up and eating food from a container is used to insert food into the mouth.*
score	Метрики от модели распознавания изображений, используемые для фильтрации и ранжирования	0.49

Используемые методы

Direct Preference Optimization

Первым используемым алгоритмом является Direct Preference Optimization(DPO)[1]

$$\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = \\ -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))\left(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\right)\right], \quad (1)$$

где $\hat{r}_\theta(x, y) = \beta \log \dfrac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ представляет вознаграждение, неявно заданное языковой моделью π_θ и референсной моделью π_{ref}.

В нашей задаче y_w и y_l такие, что $R(IoU, score|y_w) > R(IoU, score|y_l)$, где $R(IoU, score|y)$ некая функция от выводов модели детекции, отражающая пригодность промпта y для этой модели

[1] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Fin. Direct preference optimization: Your language model is secretly a reward model

Используемые методы

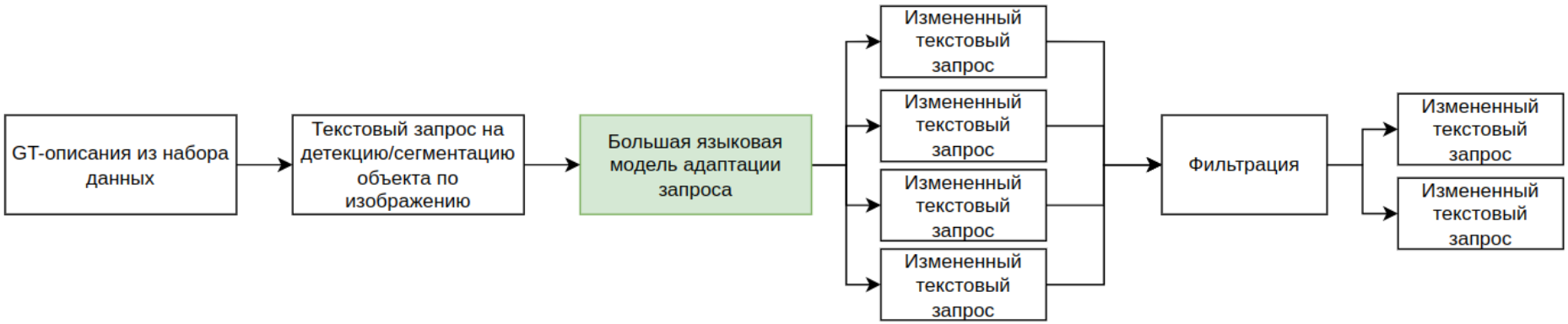
Direct Preference Optimization

Подход состоит из трех шагов:

- Аугментация данных
- Оценка и формирование датасета
- Обучение модели

Direct Preference Optimization

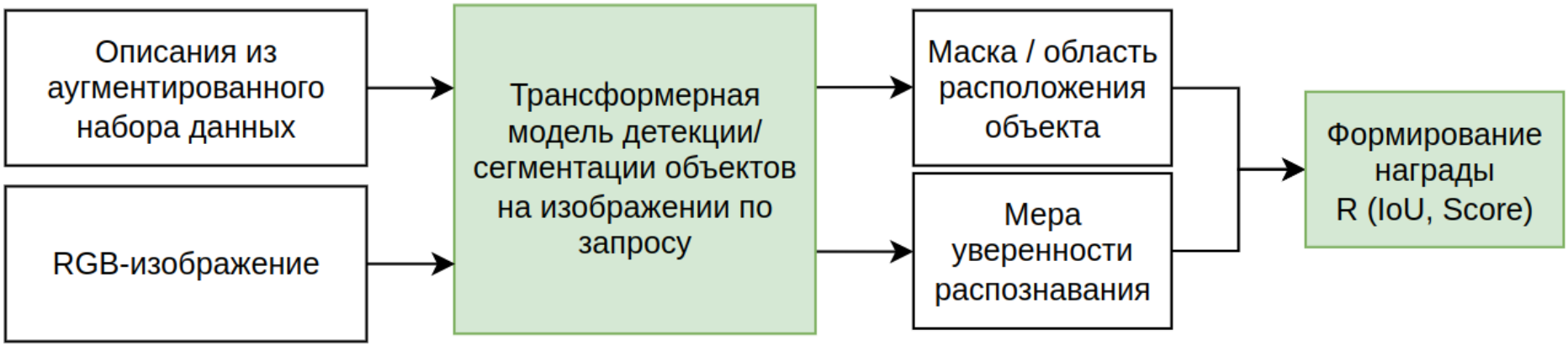
Аугментация данных



Процедура аугментации GT-данных

Direct Preference Optimization

Оценка данных и генерация датасета



Процедура оценки аугментированного набора данных

Итоговый датасет в формате DPO имеет вид:

prompt	chosen	rejected

Используемые методы

Proximal Policy Optimization

Также к данной задаче применяется алгоритм Proximal Policy Optimization(PPO)[1]

$$\mathcal{L}_{\text{PPO}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{ref}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{ref}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right], \tag{2}$$

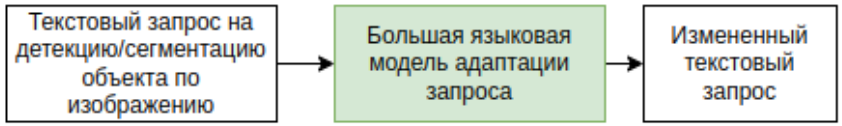
где $\hat{A}_t = \sum_{l=0}^{\infty} \gamma^l r_{t+1} - V(s_t)$, $r_t = R(IoU, score|s_t)$.

[1] John Schulman et al.: Proximal Policy Optimization Algorithms

Используемые методы

Proximal Policy Optimization

1. Обучение с учителем



2. Обучение с подкреплением

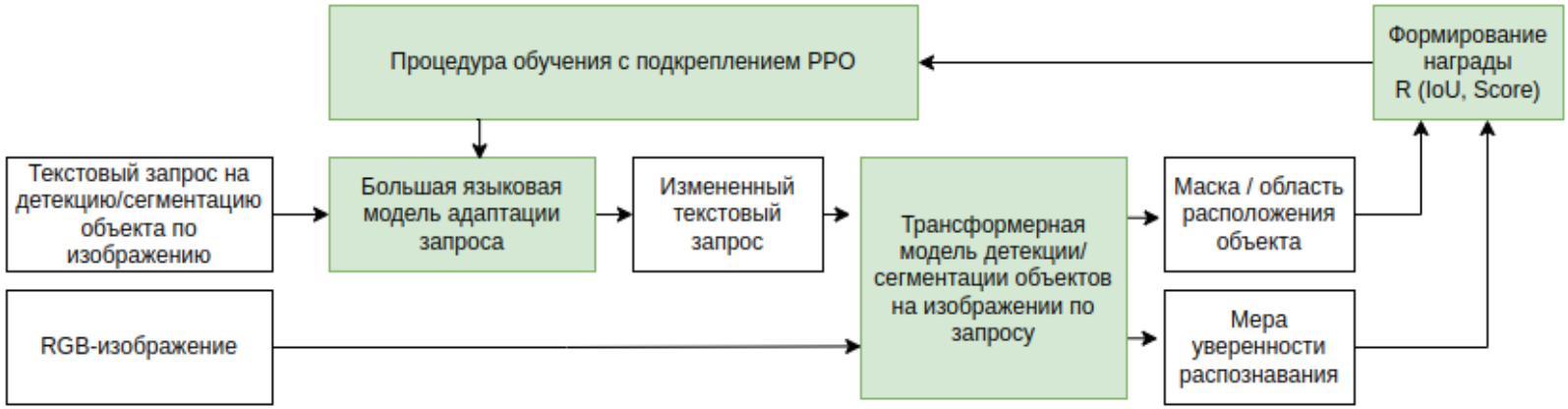


Схема предлагаемого подхода

Эксперименты

Языковые модели и модели детекции

Были проведены эксперименты с языковыми моделями с различными архитектурами и количеством параметров, а также различными моделями детекции. Были использованы языковые модели:

- Llama2[1] (7млрд параметров)
- Gemma[2] (2 млрд параметров)
- Gemma (7 млрд параметров)

Были использованы модели детекции:

- GroundingDINO
- YOLO-World

[1] Touvron, H., Martin, L., et al.: Llama 2: Open foundation and fine-tuned chat models.(2023)

[2] Gemma Team, T.M., Hardin, C., et al.: Gemma (2024)

Эксперименты

Дополнительные исследования

Также проведено несколько дополнительных исследований:

- Необходимость использования целевой модели распознавания изображений при обучении
- Эксперименты с различными вариантами функций вознаграждения для PPO:
  - Соответствующая оригинальной статье:
    $R(IoU, score) = (IoU(y) - IOU(x)) + (score(y) - score(x))$,
    где x, y - промпт и ответ модели соответственно
  - $R(IoU, score|y) = \frac{2}{\frac{1}{IoU(y)} + \frac{1}{score(y)}}$
  - $R(IoU, score|y) = IoU(y)$
- Анализ способов преобразования входных запросов

Эксперименты

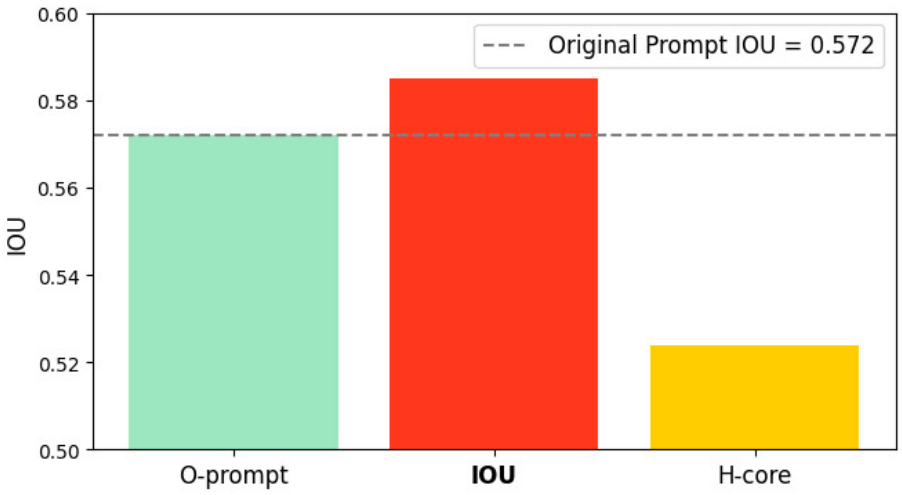
Результаты

Сравнение результатов обучения различных LLM

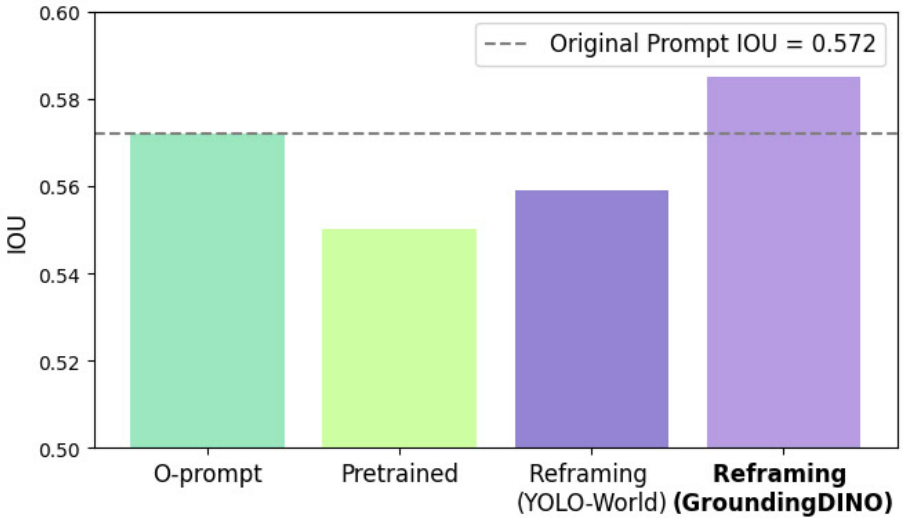
Модель распознавания изображений	**Метод обучения языковой модели**	**External, IoU**	**GOLD-A, IoU**
GroundingDINO	*Оригинальный промпт*	0.561	0.572
	Gemma-7b (Предобученная)	0.379	0.436
	Gemma-7b **DPO**	0.398	0.523
	Gemma-7b **PPO**	0.542	0.564
	Llama2-7b (Предобученная)	0.428	0.550
	Llama2-7b **DPO**	0.448	0.562
	Llama2-7b **PPO**	**0.572**	**0.585**
YOLO-World	*Оригинальный промпт*	0.501	0.535
	Gemma-7b (Предобученная)	0.270	0.433
	Gemma-7b **DPO**	0.302	0.508
	Gemma-7b **PPO**	0.442	0.521
	Llama2-7b (Предобученная)	0.411	0.511
	Llama2-7b **DPO**	0.453	0.528
	Llama2-7b **PPO**	**0.504**	**0.545**

Эксперименты

Результаты



(a)



(b)

a)Сравнение различных стратегий оценки
b)Сравнение результатов Llama2-7b, обученной с использованием различных моделей распознавания изображений, и её инференса на GroundingDINO

Эксперименты

Результаты

Примеры текстовых описаний до и после применения метода

Исходное описание	**IoU**	**Описание после применения метода**	**IoU**
A sleek, black **photo frame** with squared-off edges is standing upright on a flat surface, its imposing presence drawing attention.	0.0	This is a black **picture frame**. The frame has a sleek, modern design that looks great	0.97
The **plate** being referred to **is red in color**	0.80	The **item is a red dinner plate** that exudes a vintage aesthetic, characterized by its classic design and timeless appeal.	0.82
A *set* of **blue scissors** with handles.	0.91	A *pair* of scissors with **blue handles**.	0.93

Публикация

Результаты проведенных исследований опубликованы на международной конференции **HAIS2024**,[1] а также в серии книг **LNCS**[2]

[1] https://haisconference.eu/

[2] https://link.springer.com/chapter/10.1007/978-3-031-74186-9_11

Разработка быстродействующего метода распознавания изображений по визуальному запросу

Мотивация

- Традиционные методы распознавания изображений часто сталкиваются с трудностями при точной идентификации объектов, особенно если они сняты под неудобным углом, частично скрыты или находятся в тени.
- Использование визуальных запросов помогает решить эту проблему, однако большинство существующих решений также несовершенны: имеют проблемы с быстродействием или с качеством работы.

Постановка задачи

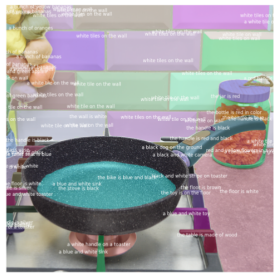
- **Входные данные**:
  - I: входное изображение, представленное как тензор размера $H \times W \times C$, где H — высота, W — ширина, C — количество каналов (например, RGB).
  - Q: визуальный запрос, который может быть представлен в текстовой форме или в виде другого изображения. В текстовом виде Q — строка или тензор, в визуальном виде Q — тензор размера $H_q \times W_q \times C_q$.
- **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.

Формально, задача состоит в нахождении отображения:

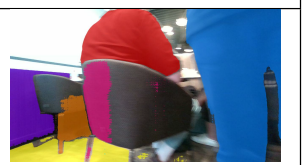
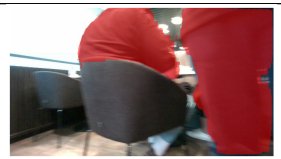
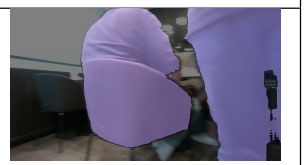
$$f : (I, Q) \underset{L \to 0}{\mapsto} M$$

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

Анализ существующих решений

Название модели	Поддерживаемые виды запросов	Быстродействие на изображении 720x1280 для RTX 2080 TI	Пример работы модели
Segment And Caption Anything	• Panoptic • Маска	2000 мс (LLM: Llama3-8B) (Nvidia H100)	
MeViS_Track_Solution Работает с видео	• Текстовый запрос	288 мс (SWIN-L)	
UniVS Работает с видео	• Визуальный запрос • Паноптик • Маска • Текстовый запрос	288 мс (ResNet50x4, SWIN-B) (Tesla V100)	
SEEM	• Panoptic • Текстовый запрос • Аудио запрос • Визуальный запрос • Маска • Различные комбинации	191 мс (Focal-L)	

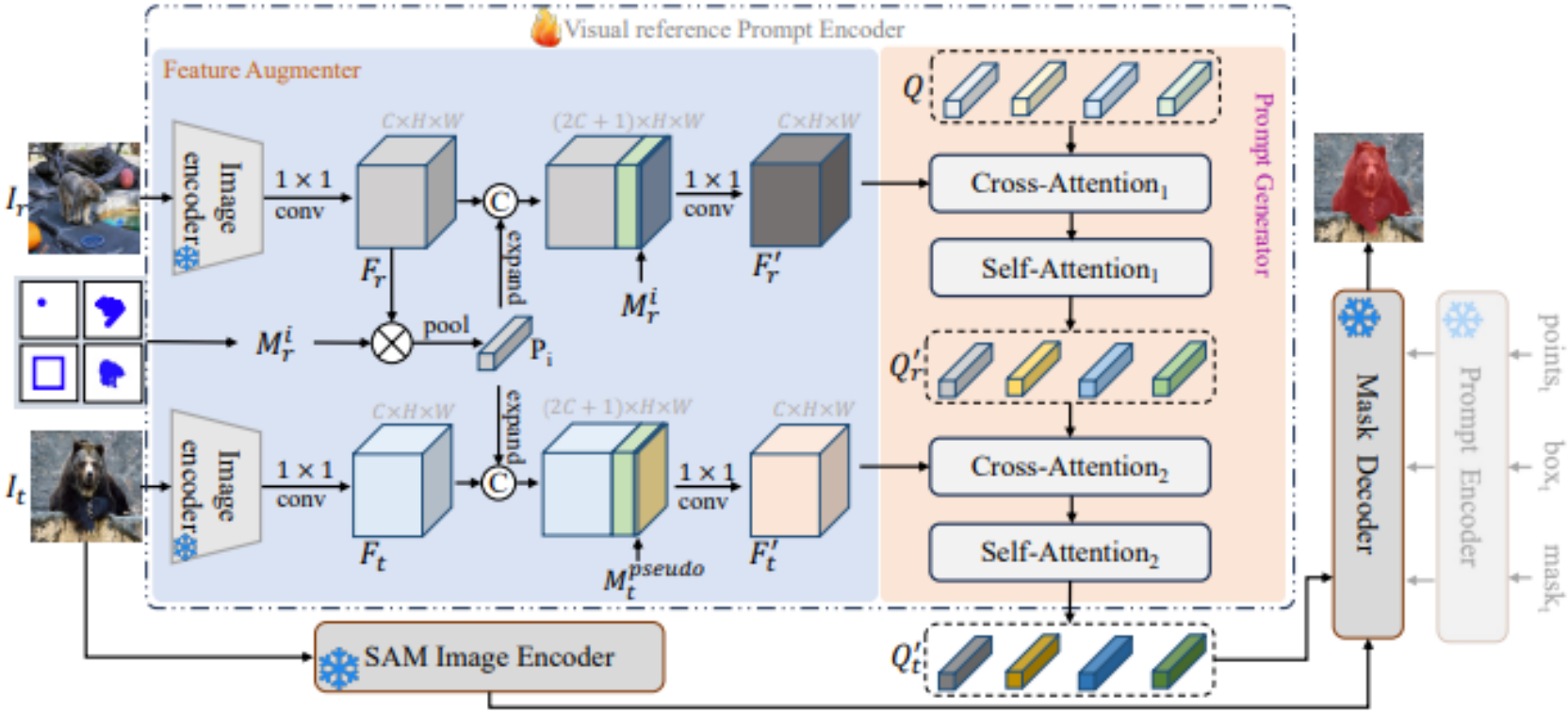
Анализ существующих решений

Название модели	Поддерживаемые виды запросов	Быстродействие на изображении 720x1280 для RTX 2080 TI	Пример работы модели
MobileSAMv2	• Клики	**75 мс** (EfficientVIT-L)	
OMG-Seg Работает с видео	• Panoptic • Маска	246 мс (ConvNeXt-L, Conv-L)	
SAM	• Клики • Боксы	540мс (VIT-H)	
SAM 2 Работает с видео	• Клики • Боксы	310 мс (1 объект) 450 мс (4 объекта) (Hiera-L) (Tesla V100)	
CLIPSeg	• Текстовый запрос • Визуальный запрос	100мс (VIT-B)	
VRP-SAM	• Визуальный запрос	680мс (ResNet101, VIT-H)	

Бейзлайн

Модель VRP-SAM

В качестве бейзлайна выбрана модель VRP-SAM:



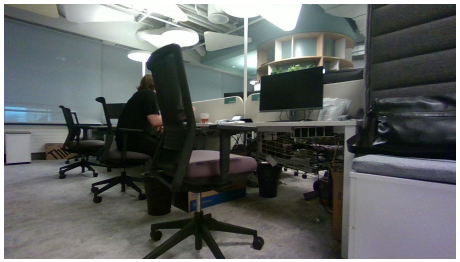
Архитектура модели VRP-SAM

Эксперименты

Проведены следующие эксперименты с бейзлайн моделью VRP-SAM:

- Эксперименты с базовыми моделями извлечения признаков для VRP-SAM
- Эксперименты с дообучением VRP-SAM

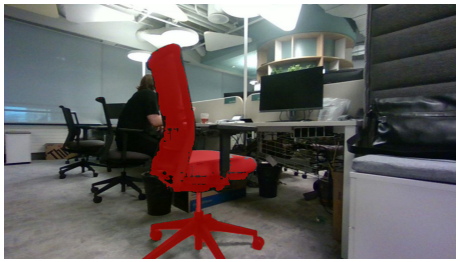
Эксперименты



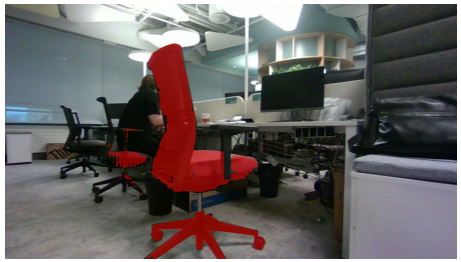
(a) Изображение-запрос



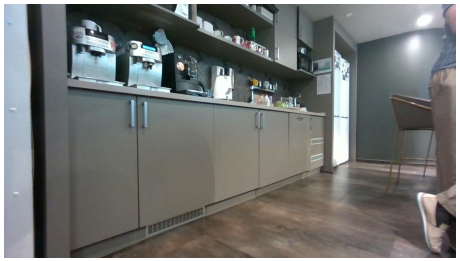
(b) Изображение-промпт



(c) VRP-SAM (ResNet101)



(d) VRP-SAM (ResNet50)



(e) Изображение-запрос



(f) Изображение-промпт



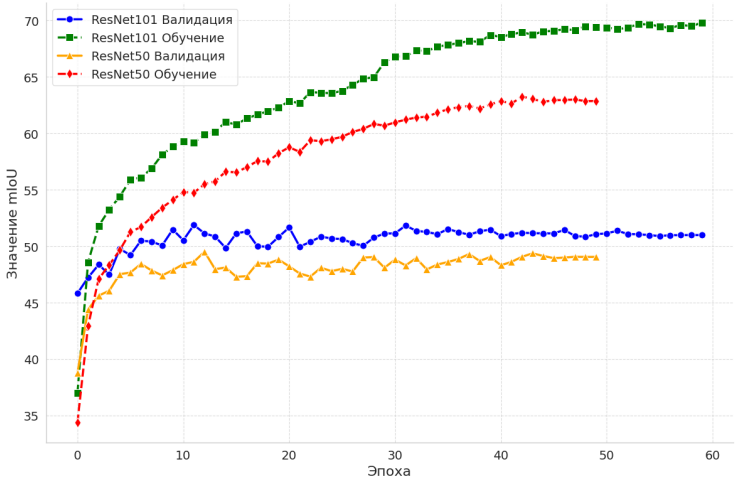
(g) VRP-SAM (ResNet101)



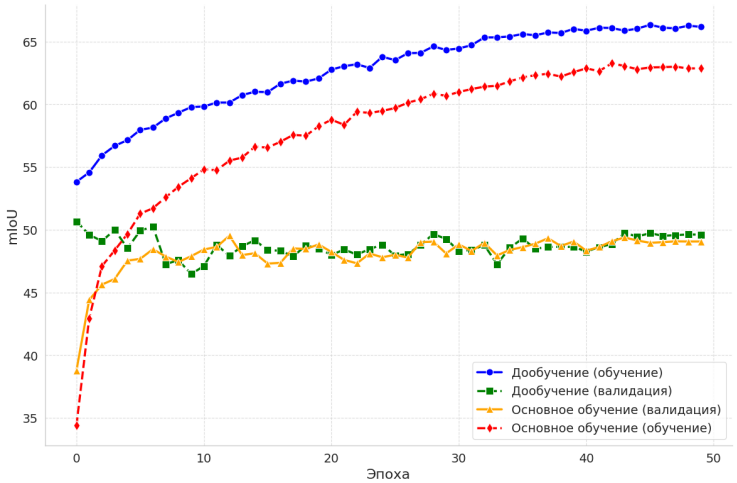
(h) VRP-SAM (ResNet50)

Сравнение качества работы моделей VRP-SAM с использованием базовых моделей извлечения признаков ResNet50 и ResNet101

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

Дальнейшие планы

- Эксперименты с моделью кодирования и предсказания масок VRP-SAM
- Продолжение экспериментов с базовыми моделями извлечения признаков
- Интеграция текстовых запросов в базовую модель VRP-SAM

Спасибо за внимание!

avshalumov.ms@phystech.edu