Сегментация объектов сложной формы по текстовому запросу

Ветошкин Л.Н.
vetoshkin.ln@phystech.edu

Центр когнитивного моделирования МФТИ

2024 – Научно-исследовательский семинар



План доклада

- Мотивация
- Постановка задачи
- Описание метода
- Проделанная работа
- Дальнейшие исследования

Мотивация

Сегментация объектов по запросу — это перспективное направление, которое применяется в различных областях, таких как робототехника, беспилотные автомобили и разметка данных. Данный тип сегментации способен сегментировать объекты различных классов.

Проблемой на данный момент является то, что современные архитектуры показывают низкое качество при сегментации объектов сложной формы.



.

Задача

Если в структуре объекта будут присутствовать тонкие линии, такие объекты будем называть - *объектами сложной формы*. Например: провода, велосипед, клетка и т.п.

Задача состоит в разработке метода, который будет максимизировать mIoU для объектов сложной формы. Формально задачу можно описать следующим образом:

$$\mathcal{C} = \{c_0, c_1, ...\} - \text{классы объектов сложной формы}$$

$$\text{mIoU}(\mathbf{M}^{\mathcal{C}}, \Phi(\mathbf{I}^{\mathcal{C}}, \mathbf{T})) \to max$$

где $\mathbf{I}^{\mathcal{C}}$ - изображения, содержащие объекты $c \in \mathcal{C}$, $\mathbf{T}$ - текстовый запрос, $\mathbf{M}^{\mathcal{C}}$ - истинные маски объектов $c \in \mathcal{C}$ находящихся на изображениях $\mathbf{I}^{\mathcal{C}}$, Φ - разработанный метод.

Проделанная работа

- Проведено обучение моделей HQ-SAM и модели Text-HQ-SAM на датасете HQSEG-44K[1]
- Произведена оценка моделей
- Были выявлены недостатки Text-HQ-SAM
- Разработаны новые методы для устранения выявленных недостатков
- Собран новый датасет для обучения модернезированной модели

[1] Ke L. et al. Segment anything in high quality //Advances in Neural Information Processing Systems. – 2024. – Т. 36.

Результаты

Результаты на датасете TAOMR[1]:

Модель	bump	manhole	pit	puddle	firehose	hose	wire	waste	mean
Text-HQ-SAM	**0.945**	**0.795**	0.895	0.838	**0.887**	**0.663**	**0.760**	**0.883**	**0.833**
HQ-SAM (trained)	0.943	0.794	**0.904**	**0.843**	0.860	0.654	0.750	0.872	0.828
HQ-SAM (pretrained)	0.929	0.724	0.813	0.793	0.790	0.468	0.647	0.863	0.753

Результаты на датасете Dis5k[2]:

Модель	IoU
Text-HQ-SAM	**0.777**
HQ-SAM (trained)	0.770

Результаты на датасете ThinObject-5K[3]:

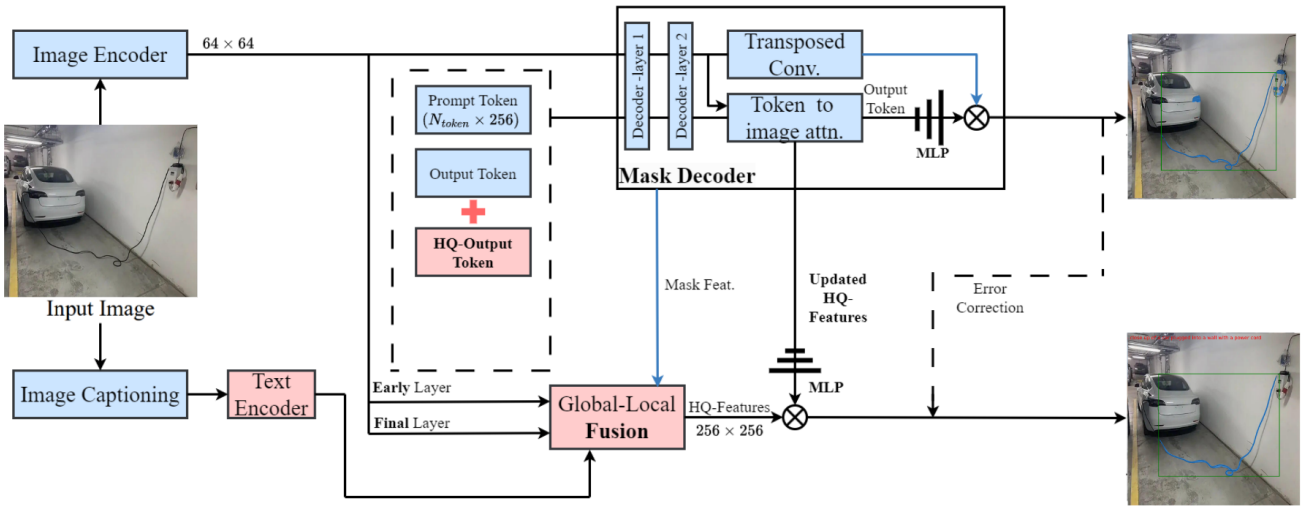
Модель	mIoU
Text-HQ-SAM	0.817
HQ-SAM (trained)	**0.822**

[1]Vetoshkin, L., Matykina, O., Kichik, M., Izmesteva, U., Yudin, D. How to Segment Wires, Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot. Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot.

[2]Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In ECCV, 2022

[3]Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In WACV, 2021

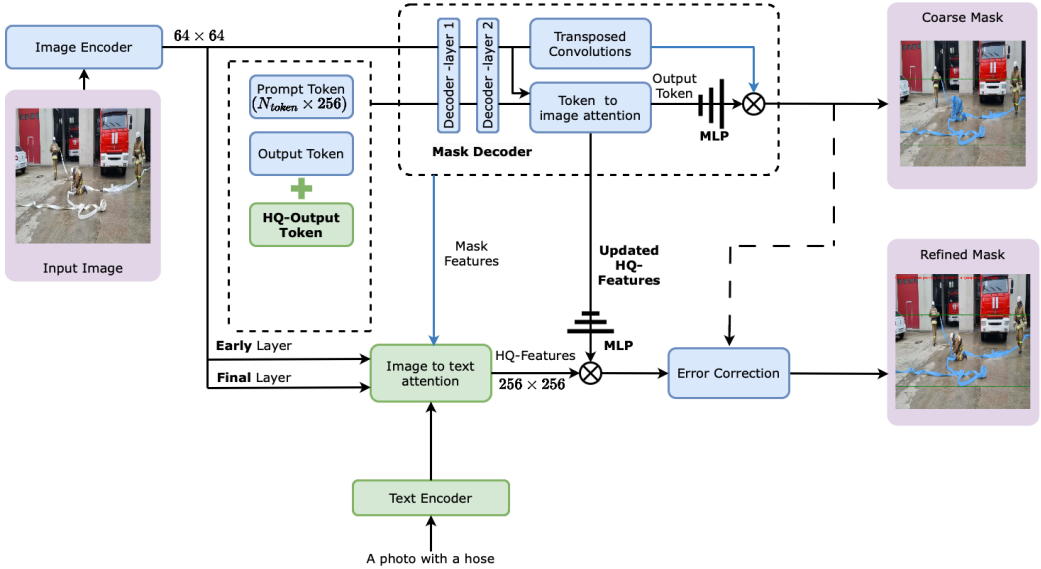
Text-HQ-SAM



Данная модель обладает следующими недостатками:

- Генерация описания к изображению часто является неточной
- Блок слияния признаков (Global-Local Fusion) плохо извлекает семантическую связь, так как представляет собой просто суммирование признаков.

Text-HQ-SAMv2



- Удалён модуль Image Captioning, так как сгенерированные описания часто не соответствовали содержанию изображения
- Сформирован собственный датасет, включающий объекты с тонкими линиями из различных областей (медицина, робототехника, инженерия и т.д.). Аннотации представлены в виде текстовых описаний и масок
- Добавлен contrastive loss для сопоставления признаков текста и визуальных признаков, извлечённых из области предсказанной маски

Image to text attention

- Вместо суммирования признаков используется cross-modality multi-head attention module (X-MHA), аналогично GLIP:

$$O = \text{Enc}_I(\text{Img}), \quad P = \text{Enc}_L(\text{Prompt})$$

$$O^{(q)} = OW^{(q,I)}, \quad P^{(q)} = PW^{(q,L)}, \quad Attn = \frac{O^{(q)}(P^{(q)})^\top}{\sqrt{d}},$$
$$P^{(v)} = PW^{(v,L)}, \quad O_{\text{t2i}} = \text{SoftMax}(Attn)P^{(v)}W^{(out,I)},$$
$$O^{(v)} = OW^{(v,I)}, \quad P_{\text{i2t}} = \text{SoftMax}(Attn^\top)O^{(v)}W^{(out,L)}.$$

$$O^i_{\text{t2i}}, P^i_{\text{i2t}} = \text{X-MHA}(O^i, P^i)$$

Датасет

Датасет состоит из изображений содержащих объекты состоящих из тонких линий. Текстовые описания были сгенерированы с помощью модели BLIPv2.
Благодаря знанию о том, какой объект находится на изображении, текстовые запросы для BLIPv2 генерируются по шаблону:

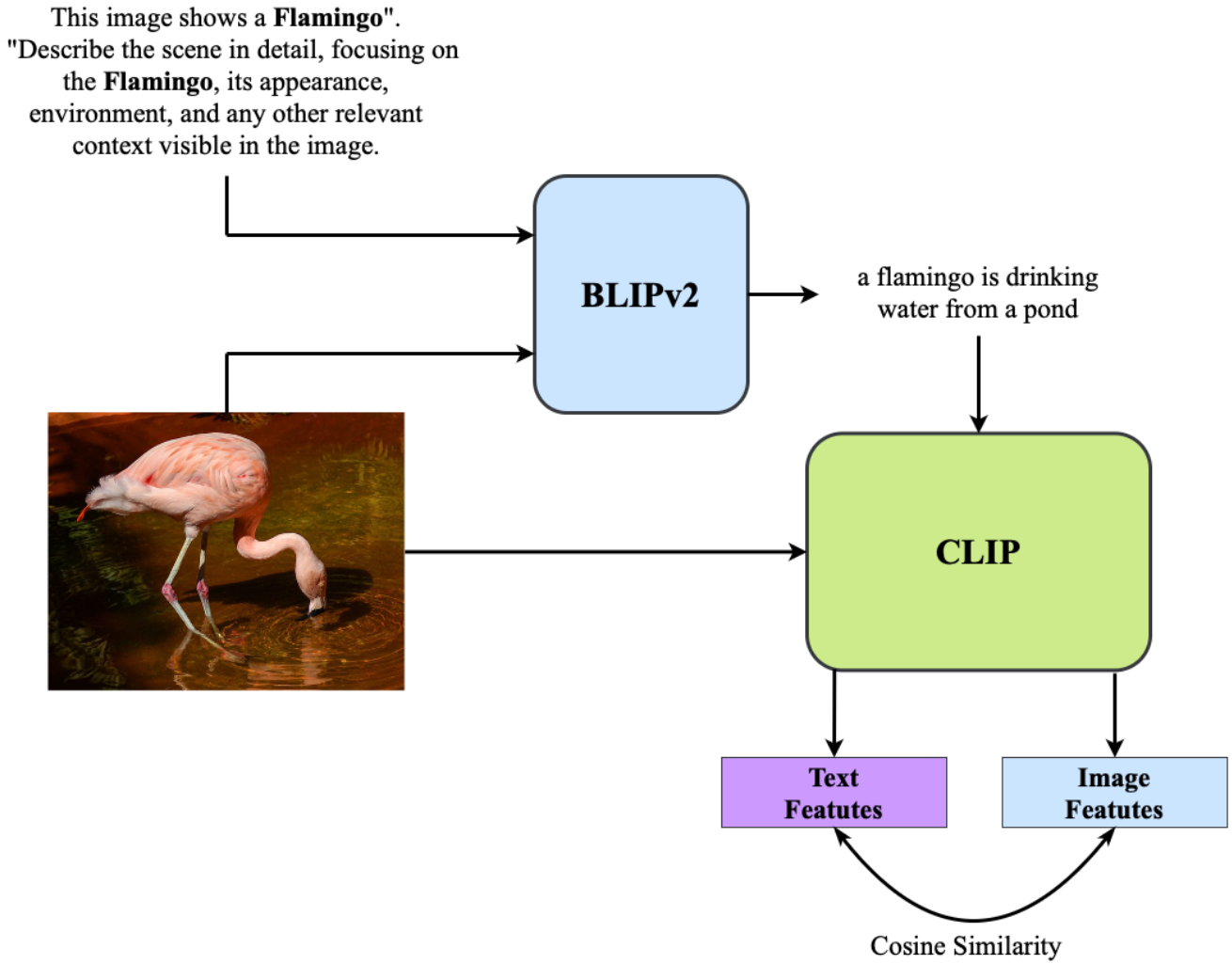
"This image shows a [CLS].Describe the scene in detail, focusing on the [CLS], its appearance, environment, and any other relevant context visible in the image."

Также с помощью модели CLIP оценивается косинусное расстояние между признаками сгенерированного текста и признаками изображения. Если расстояние мало, то описание генерируется по шаблону:

"A photo with a [CLS]"

Датасет

Ниже показана схема для генерации текстовых аннотаций к изображениям:



Дальнейшие исследования

- Обучить модели Text-HQ-SAMv2 и HQ-SAM на новом собранном датасете и оценить их качество с использованием различных метрик
- Провести обучение с использованием различных текстовых описаний для анализа их влияния на результаты сегментации
- Выполнить эксперименты с разными функциями потерь для выбора оптимальной
- Измерить время работы Text-HQ-SAMv2 и сравнить его с другими моделями
- Исследовать различные методы уточнения маски и их влияние на итоговое качество сегментации

Спасибо за внимание!

cogmodel.mipt.ru

vetoshkin.ln@phystech.edu