CLIP-Driven Approach on Part-Granularity Referring Image Segmentation

Mo Zhou
University of Colorado Boulder

1. Introduction

Object part segmentation is an important task in computer vision, because it involves identifying and segmenting specific parts of an object, rather than just classifying the object as a whole. This fine-grained understanding of objects is crucial in various fields. For robots to effectively interact with expected objects, understanding parts will help them to decide where to grasp, manipulate or perform more delicate actions. In augmented reality, understanding parts separately allows user to interact with objects more flexibly, such as rotating wheels or posing toys.

Meanwhile, Datasets is always one of the biggest challenges for research in computer vision, since most computer vision tasks prefer labeled datasets, and for tasks in different area, datasets of specialized types of images are also needed. This leads to many limitations of previous models. For a long time, the research has been focusing on constructing large-scale datasets. This is reasonable, because increasing the amount of training data is the most direct way of improving model performance. However, people did not pay equal attention to the information density of the dataset, which means the information each image contains. Models trained on these large-scale datasets perform well on coarse-grained tasks, but suffer from fine-grained tasks. One example is that they can hardly recognize a certain part of an object. Recently, some datasets came out to fill this blank. PartImageNet[4] provides a large, high-quality dataset with part segmentation annotations, consisting of 158 classes from ImageNet with approximately 24,000 images. Going deeper, SPIN[7] adds subpart annotations of objects on top of that. With part-level annotations, models are able to learn the image composition in finer granularity. Unfortunately, because of the fact that these annotation works inevitably require human participation which is expensive, these existed part-level annotation datasets are still in relatively small scale. To overcome this shortage on available training data, we need to find a new way to improve the efficiency of learning.

In recent years, large language models have been proven to be powerful on solving all kinds of downstream tasks, thanks to not only the transformer architecture with attention mechanism, but also the ability that it can learn from abundant on-the-shelf raw text on the Internet. Text data on the Internet is uploaded by people in different industries from all over the world, so large language model trained on that exhibits extraordinary generalization ability and scalability. Inspired by that, CLIP[8] has successfully replanted this idea into computer vision tasks. It encodes images and texts into a common latent space, where the matched text and image should be close to each other. This kind of data is much easier to access, so the pre-trained model has great generalization ability on a lot of downstream tasks. Recent studies on CLIP leverage its output representation on transformer architectures to efficiently transfer to popular computer vision tasks[1] and it turns out to be a good way to enrich the knowledge base of models while keeping training costs practical.

Yet, the application of CLIP on part segmentation tasks is still under explored. In this paper, we propose a CLIP-driven approach with fine-tuning on PartImageNet to achieve better performance with limited data resources. This will retain the generalization ability on open-vocabulary classes from CLIP, while further enhancing the sensitivity on object parts.

2. Related Work

Part-level datasets After Imagenet[3] came out, research on image datasets becomes focusing more on constructing large-scale datasets[5, 10]. With sufficient data, the model is more likely to learn diverse information, and more parameters could be used without worrying about overfitting. While large-scale datasets are preferred for many tasks, most of them did not provide fine-grained annotations on object parts. Some research has come out to fill this blank[2, 9]. However, there is still a lot of information missing considering various classes of parts, so we choose to use CLIP as an efficient method of learning from natural language supervision to leverage large amounts of publicly available data on the internet as the knowledge base.

CLIP on downstream tasks After CLIP becomes popular, many studies focus on exploring its potential in different computer vision tasks. CLIP-Decoder[1] proposed a novel way by combining multimodal representation encoded by CLIP with transformer architecture to achieve zero-shot

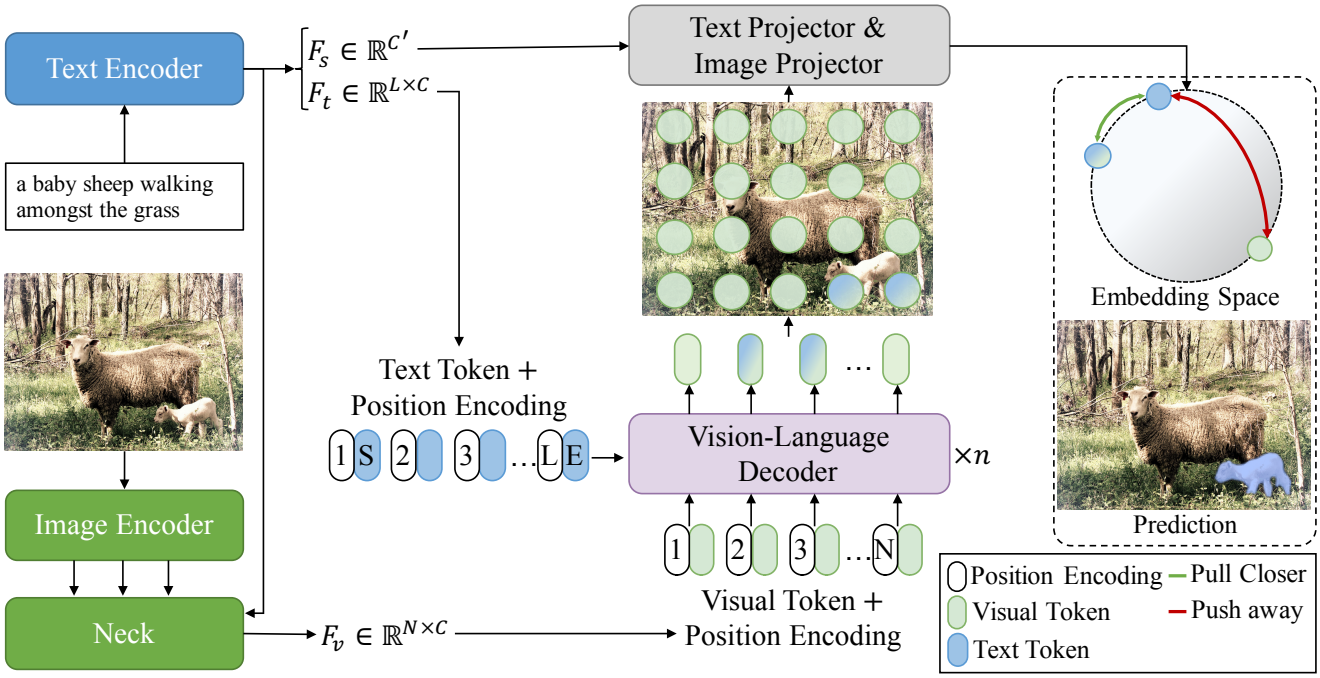

Figure 1. **The overview of CRIS framework from their paper.** Besiders original CLIP structure, it adds on a cross-modal neck, a vision-language decoder, and two projectors. The pixel-level information is learnt by the novel text-to-pixel contrastive learning loss. **Note: The example input image given in the figure is for segmenting an object, whereas we will focus on segment parts such as the feet of the baby sheep.**

Figure 2. An example of annotated images in PartImageNet. Parts are segmented based on its superclass.

multilabel classification. Yuqi et al.[6] made improvements on CLIP to localize different categories with only image-level labels and without further training. However, none of these works are meant to solve part-level object segmentation, which makes our work valuable.

Referring Image Segmentation The goal of referring image segmentation is to segment an object from an image based on a natural language expression, called a referring expression. The model must interpret the referring expression, find the relevant object in the image, and output a pixel-wise segmentation mask. This differs from traditional segmentation tasks because it involves understanding and linking a specific language query to visual content.

Now with the knowledge of CLIP, Zhaoqing et al.[11] proposed a framework that can understand multi-modal information and match cross-modal features better. It refines pixel-level visual features with textual features, and do text-to-pixel contrastive learning to offer the model more fine-grained information. However, their focus stays on object-level tasks. Their training dataset and benchmarks do not contain enough part-level annotations. Therefore, in our paper, we implement CRIS as the backbone model, with fine-tuning on part annotation dataset, to explore its potential on segmenting part instances.

3. Methods

To perform a segmentation task with CLIP-driven architecture, we use CRIS as the backbone model. The detailed architecture is shown in figure 1, and we will discuss the key parts in the following sections.

Image Encoder The image encoder of CRIS takes as inputs images $M \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the height and width of the image. We follow their settings to use ResNet to extract visual features from different stages. Throughout the fine-tuning on PartImageNet, the image encoder will learn knowledge about how to define object parts.

Text Encoder The textual features $F_t \in \mathbb{R}^{L \times C}$ are ex-

tracted by a Transformer from the input $T \in \mathbb{R}^L$ which serves as a referring expression, i.e., a part name you want to segment out. With the self-attention mechanism, the encoder will capture the relationships between words. The input text representation is generated by lower-case Byte Pair Encoding(BPE), bracketed with [SOS] and [EOS] tokens. In the final layer, the embedding of [EOS] token will be further transformed as a global textual representation $F_s \in \mathbb{R}^{C'}$. L is the length of input text, while C and C' are the feature dimensions.

Vision-Language Decoder To better combine textual features into visual features, in CRIS a novel decoder module including multi-head self-attention and cross-attention layers is introduced. It takes as inputs the textual features $F_t \in \mathbb{R}^{L\times C}$ and pixel-level visual features $F_v \in \mathbb{R}^{N\times C}$ and outputs a sequence of multi-modal features $F_c \in \mathbb{R}^{N\times C}$. Due to the unawareness of positional information of transformer architecture, the positional embeddings are added on F_t and F_v respectively.

Text-to-Pixel Contrastive Learning Loss To capture the alignment between textual description and pixel-level information, CRIS use text-to-pixel contrastive learning. The global textual features F_s and decoded multi-modal features F_c will first go into two projectors to get Z_v and Z_t, and then we will do contrastive learning on them based on the text-to-pixel contrastive loss that can be formulated as:

$$L^i_{con}(z_t,\, z^i_v) = \begin{cases} -\log \sigma(z_t \cdot z^i_v), & i \in \mathcal{P}, \\ -\log(1-\sigma(z_t \cdot z^i_v)), & i \in \mathcal{N}, \end{cases} \quad (1)$$

$$L_{con}(z_t,\, z_v) = \frac{1}{|\mathcal{P}\cup\mathcal{N}|} \sum_{i\in\mathcal{P}\cup\mathcal{N}} L^i_{con}(z_t,\, z^i_v), \quad (2)$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of "1" and "0" in the ground truth, $|\mathcal{P}\cup\mathcal{N}|$ is the cardinality, σ is the sigmoid function. The final segmentation is made by reshaping $\sigma(z_t \cdot z_v)$ into $\frac{H}{4}\times\frac{W}{4}$ and upsampling it back to the original image size.

4. Experiments

To inject the knowledge of object parts, the proposed approach is fine-tuned on PartImageNet train split and compared with popular solutions on object segmentation. We evaluate its effectiveness on PartImageNet test split by IoU and Precision@X.

4.1. Dataset

PartImageNet is a large, high-quality dataset with part-level annotations based on ImageNet. In total 158 classes from ImageNet was selected and being grouped into 11 superclasses following the WordNet hierarchy. Parts are labeled according to these superclasses. The available number of images in PartImageNet is about 24k, which makes it large enough for transfer learning.

We train our model on PartImageNet train split containing 20,481 images and 95,059 parts, and do evaluation on test split which has 2,408 images and 11,275 parts.

4.2. Experimental Settings

We follow the original CRIS settings with input image size to be 416 $\times$ 416 and the number of transformer decoder attention heads to be 8. The number of epochs for fine-tuning is planned to be 3-10 with Adam optimizer and learning rate λ = 0.0001.

To prove the efficacy of the proposed approach on part segmentation tasks, we compare its performance with Mask R-CNN. We will use the pre-trained Mask R-CNN as one baseline model, and then we will also do the same fine-tuning on it as another baseline to figure out if such a CLIP-driven approach with fine-tuning is more powerful than a routine segmentation model.

4.3. Metrics

To evaluate the effectiveness of our model, we use two common metrics in object segmentation: IoU and Precision@X.

The IoU (Intersection over Union) calculates the ratio of the intersection area to the union area between the predicted segmentation mask and the ground truth mask. It provides a direct measure of how well the predicted mask overlaps with the actual object.

Precision@X measures the percentage of test images where the IoU score exceeds a given threshold $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. This metric assesses the localization ability of the model.

References

[1] Muhammad Ali and Salman Khan. Clip-decoder : Zeroshot multilabel classification using multimodal clip aligned representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4675–4679, October 2023.

[2] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[4] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. *arXiv preprint arXiv:2112.00933*, 2021.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[6] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation, 2023.

[7] Josh Myers-Dean, Jarek Reynolds, Brian Price, Yifei Fan, and Danna Gurari. Spin: Hierarchical segmentation with subpart granularity in natural images, 2024.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[9] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7141–7151, June 2023.

[10] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[11] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation, 2022.