CLIP on Object Part 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 to use CLIP with fine-tuning on PartImageNet to efficiently utilize limited datasets. 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 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.

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.