Enhanced Object Proposal Generation for Weakly Supervised Instance Segmentation

Karim Umar
umarka@student.ethz.ch
ETH Zürich
Switzerland

ABSTRACT

Machine Learning has benefited immensely from scaling datasets both in terms of size and variety. On tasks where such large and diverse datasets have been applied, monumental increases in both robustness and ability to generalize have been shown. Datasets for both object detection and instance segmentation lag behind this trend. Traditionally requiring human annotations, no clear path to scaling these datasets to the size and diversity, shown to be so beneficial in other domains, has been identified.

This thesis examines an alternative approach to generating weakly supervised object detection and instance segmentation datasets. It differentiates itself from previous weakly supervised approaches by leveraging language modelling as part of its annotation pipeline to extract relevant information from image captions and thus achieve a greater variety of pseudo labels.

1 RELATED WORK

Many works [1, 2, 3] have proposed methods to leverage the abundantly available image-text datasets to train object detection and instance segmentation models. These approaches typically achieve this with a weakly supervised approach for existing image-text datasets, deriving pseudo-labels from the image's caption, either using the whole caption [1, 3] or by deriving pseudo labels from the image caption (for example by generating n-grams of the words in an image's caption [2]). These approaches suffer from the limitation that image captions typically describe an image as a whole and often omit objects visible in the image, as this is redundant to a human reading the caption.

2 APPROACH

2.1 Data Generation

We source our data from the High-Resolution subsets of LAION [4], these image-text pairs are filtered to exclude images with non-English captions and potential NSFW content.

In this work, we concentrate mainly on the object proposals. We want to mitigate the fact that image captions generally omit explicitly naming objects that are visible in an image, but whose presence would be obvious to a human and hence not be named explicitly (for example, "`leaf`" in an image of a plant, "`wheel`" in an image of a car or "`window`" in an image of a house).

To derive pseudo-label candidates, the image caption is wrapped in different prompt templates, which are fed into an LLM to produce object candidates. The LLM used is the bigscience-7B [5], used due to restrictions of the training infrastructure.[1] These candidates are used to query an open-vocabulary object detector, resulting bounding boxes and CLIP encodings of the labels are kept. We use Grounding-DINO [6] as our open-vocabulary detector.

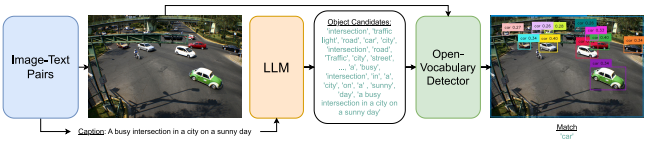

Figure 1: Data Generation Overview

2.2 Model Architecture

Semantic-SAM is based on SAM, as SAM already exhibits excellent segmentation ability. Further, the architecture of SAM, with a large image encoder and lightweight mask decoder, makes it ideal for downstream applications, especially those that allow pre computation of image features.

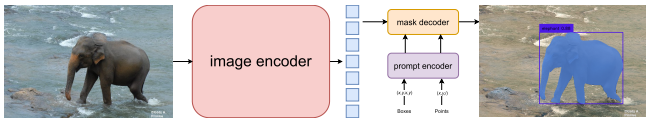

Figure 2: Model Architecture

Modifications to SAM are minimal, with the mask decoder featuring 5 extra learned embeddings, which correspond to 4 masks and 1 Cosine Similarity score predictions. Semantic-SAM outputs a semantic prediction as a 768-dimensional vector, aligned with CLIP text embeddings.

2.3 Training

Semantic-SAM is trained jointly on SA-1B [7] and our generated dataset, mixed at a ration of 1:4. This is done to maintain the excellent mask quality of SAM. For samples from SA-1B, supervision is the same as SAM training (Dice loss + Focal loss + MSE for IoU Score). For samples from our generated dataset, loss is calculated contrastively across a batch.

3 LIMITATIONS AND FUTURE WORK

This approach features numerous limitations:

(1) **Dataset Size:** We scaled our dataset to 3.9M images, constrained by available compute resources.
(2) **Bias:** The Language model and Detector used to generate pseudo annotations feature considerable bias, which will affect the quality of the generated data. Chapter 5 of the full report goes into more depth.

[1] The university cluster features many 24Gb cards, and very few 80Gb cards.

REFERENCES

[1] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, Springer, 2022, pp. 350–368.

[2] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *arXiv preprint arXiv:2306.09683*, 2023.

[3] R. Arandjelović, A. Andonian, A. Mensch, O. J. Hénaff, J.-B. Alayrac, and A. Zisserman, "Three ways to improve feature alignment for open vocabulary detection," *arXiv preprint arXiv:2303.13518*, 2023.

[4] C. Schuhmann *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[5] *bigscience/bloom-7b1 · Hugging Face*, [Online; accessed 29. Aug. 2023], Aug. 2023. [Online]. Available: https://huggingface.co/bigscience/bloom-7b1.

[6] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[7] A. Kirillov *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.