Task 2. Open-vocabulary segmentation

Task

In this task you will be exploring the capabilities of pre-trained foundation models: segment anything(https://github.com/facebookresearch/segment-anything) and CLIP(https://github.com/openai/CLIP).

Your task will be the utilization of these powerful models for object retrieval. Given the image set and textual prompt you will have to retrieve all masks of all objects of this type on the scene.

The idea is simple, in order to perform retrieval you would have to:

1. segment an image
2. find those masks that correspond to object of interest(for example based on cosine similarity between text embedding and segment embedding)
3. combine these masks together(i.e. perform logical or operation on 2 masks)

To reduce the computational requirements take every 10th image(that would give you 34 images in total)

Note: you don't have to train any image segmentation/classification network in this task. Only assess zero-shot capabilities

Data

Dataset is a part of scannet++ https://github.com/scannetpp/scannetpp(scene 0a7cc12c0e) consisting of images and masks(for evaluation only)

- folder "gt_semantic_2d" contains masks in 2 formats JPG and npy, use whatever is more convenient
- folder "images_0a7cc" contains rgb images

Load data from this link:
https://drive.google.com/file/d/1hcGkpo39lp6PuinYJctipuCedJuoLxrz/view?usp=sharing

Evaluation

Evaluation will be performed based on mean iou between your estimated masks and gt masks(70% of grade) and your report(30% of grade).

Grades:

1. Metrics(for class "chair", which has class 8 in gt masks):
   iou >= 0.5 100% of points
   0.5 > iou >= 0.4 80% of points
   < 0.4 60% of points
2. Report:

- your observations (in a couple of words: tell what you learned/observed while working on this hw, try to propose possible improvements, assess your results)
- visualization of results (iou calculation, mask visualization, failure cases)

Deliverables

1. In your submission(ipython notebook) you should show a way to run your solution
2. Additionally write your observations, do research on cases where models fail and where they succeed. Investigate how prompt, mask size, proportion of background affect the performance.