Colab

Part 1: GradCAM

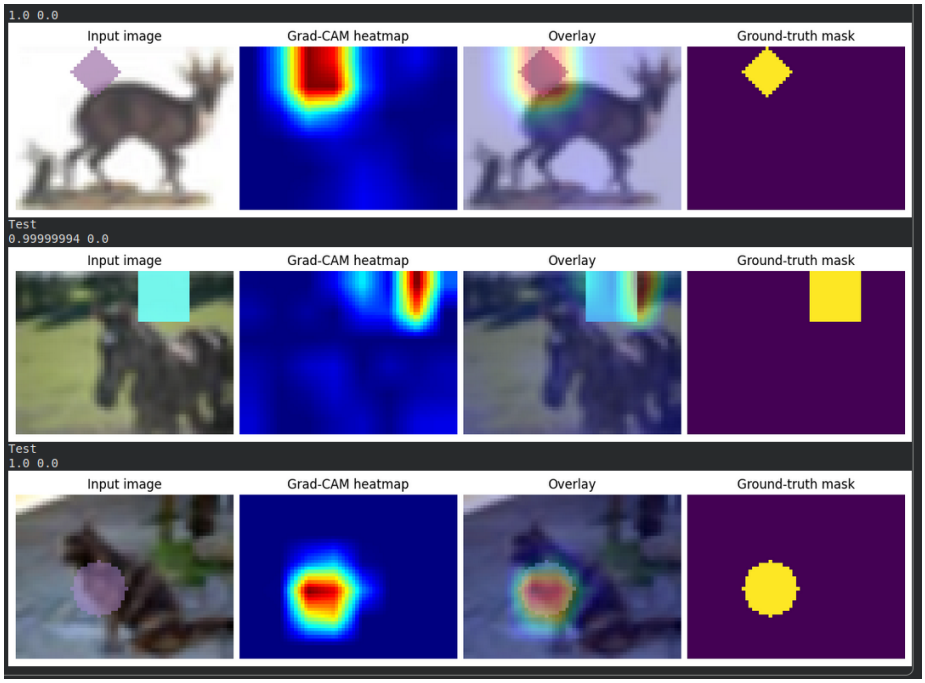
For GradCAM, we will use forward hook in order to collect the activations, and backward hook to collect the gradients.

During the model inference run, we will save the relevant activations and gradients in the class. Then we can calculate the heatmaps via a weighed product of both.

Since the target layers are not necessarily the same size as the output image, we may need to scale the heatmap to match the image size. For that we can use standard Torch interpolation.

Using `einsum`, we can do all the necessary calculations loop-free.

Results:



As we can see from the heatmap, the figure gets located correctly. In case of the square, the straight edge is the most important feature that distinguishes it from other figure types, while for a rhombus or a circle, the heatmap covers the entire figure.

Part 2: Segmentation

We will try two possible approaches to finding a suitable foreground point:

- a more simple one, with the brightest point in the heatmap serving as our foreground point;
- or one that calculated the weighed center of the heatmap after applying an initial cutoff to exclude the values irrelevant to the mask (we took 0.6).

For the target labels, we will pass no value, letting the base model infer the class on its own.

For background point, we will take the minimal value of the mask; since the area outside the mask is much larger than within the mask, it should be almost guaranteed to succeed.

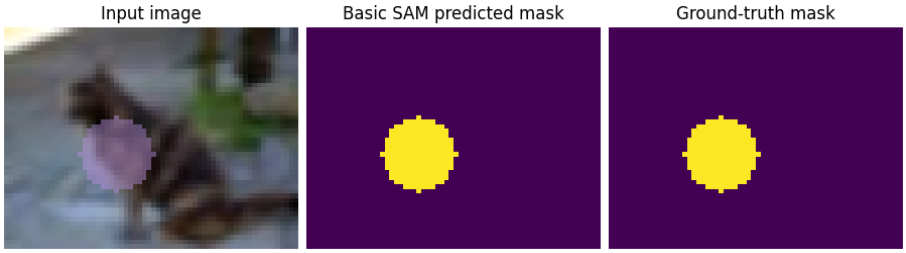
The results we get for both algorithms:

```
Hit rate (weighed centroid): tensor(0.8800)
Average distance  (weighed centroid): 4.162355788410736
Hit rate (max point): tensor(0.7660)
Average distance (max point): 5.587848964898276
```

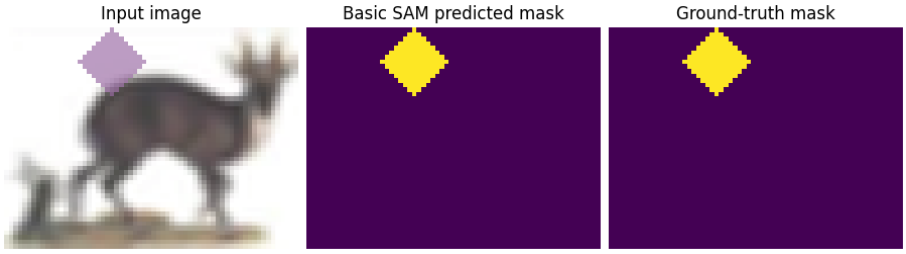
So max point algorithm gives decent results, but weighed centroid performs better; we will proceed with this approach.

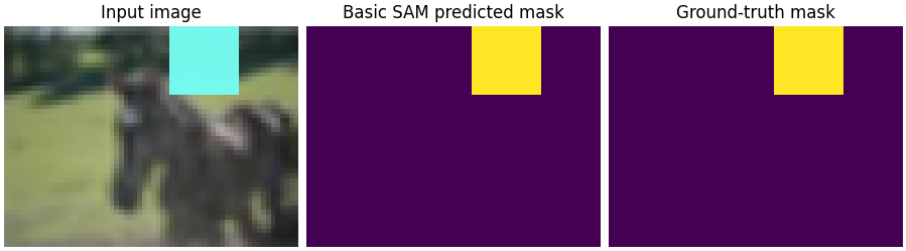
Results:

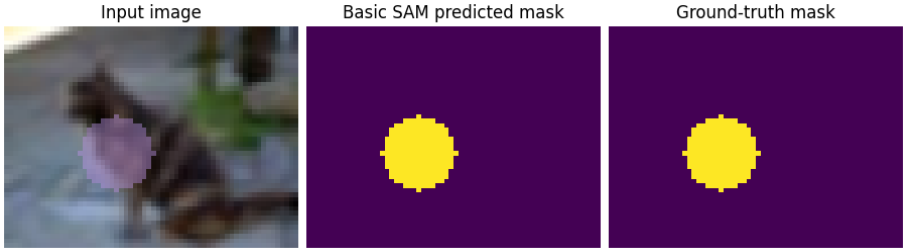
Foreground only



Foreground and background







Now if we measure IoU over all test images, we get:

`Average IoU: 0.90007424` for foreground-only approach

and

`Average IoU: 0.89502734` for foreground-background approach.

The difference is negligible.

Possible improvements

- For finding the foreground point: experiment with the cutoff from ask to non-mask. Consider making it dynamic - for example, based on modal/medial value of the heatmap.
- For finding the background point: provide not just some point outside the mask, but detect what can potentially interfere with segmentation - what objects, like the deer body in the example, are likely to be misidentified as foreground.
- For both of these, provide multiple candidate points.