Image segmentation

Deep learning - BMDC 2025-2026

Gheorghe Cosmin Silaghi

Universitatea Babeș-Bolyai

October 16, 2025

Computer vision tasks

- Image classification: assign one or more labels to an image. Could be: **single-label classification** (classes are mutually exclusive) or **multilabel classification**
- Image segmentation: the goal is to **segment** (or partition) an image into different areas, with each area representing a category
- Object detection: the goal is to draw rectangles (**bounded boxes**) around the objects of interest in an image and associate each rectangle with a class. a self-driving car uses object detection to monitor other cars, pedestrians and signs in view of its cameras

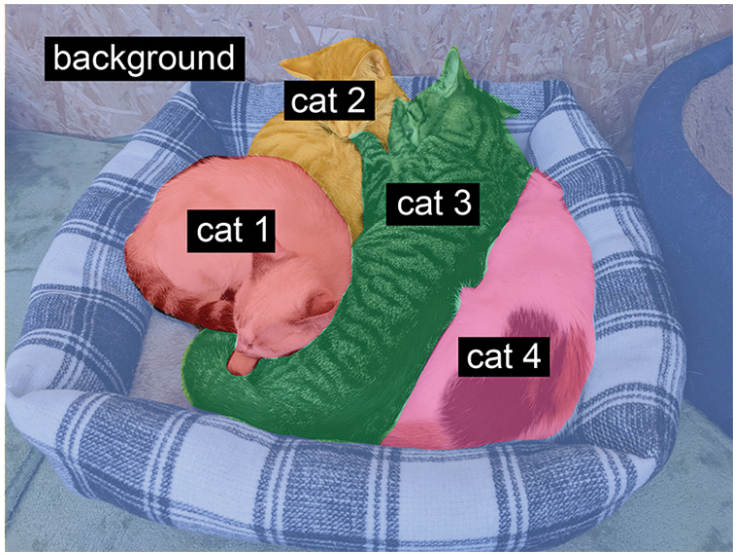
Other niche tasks

- image similarity scoring: how visually similar two images are
- keypoint detection: pinpointing attributes of interest in an image, such as facial features
- pose estimation
- 3D mesh estimation
- depth estimation etc.

Types of image segmentation

- image segmentation: using a model to assign a class to each pixel in an image. It is segmenting the image into different zones, such as background or foreground, or road, car or sidewalk

1. semantic segmentation: each pixel is classified in a semantic category, like cat
2. instance segmentation: seeks to parse out individual object instances. if two cats are in the image, segmentation will distinguish pixels belonging to cat1 and cat2.
3. panoptic segmentation: combines semantic segmentation with instance segmentation. assigns to each pixel both a semantic label (like cat) and an instance label (like cat 1).



Training a segmentation model from scratch

- segmentation mask: an image segmentation equivalent of a label: it's an image of the same size as the input with a single color channel where each integer value corresponds with a class of the corresponding pixel, like 1 (foreground), 2 (background) and 3 (contour)
- The stack of Conv2D layers downsample the image, while gradually increasing filter size
- We use strides=2 instead of MaxPooling for downsampling, because in we care about the spatial location of the information in the image. With MaxPooling the spatial location information is being destroyed.
- Second half of the model is composed of Conv2DTranspose layers that upsample the feature map (the inverse operation to Conv2D).

Intersection over Union (IoU) metric

- measures the match between the ground truth segmentation masks and the predicted masks

1. compute the intersection between the masks, i.e. the area where prediction and ground truth overlaps
2. compute the union of the masks, i.e. total area covered by both masks combined. This is the whole space we are interested in
3. Divide the intersection area with the Union area and get IoU. 1 denotes a perfect match, 0 denotes a complete miss

Using a pre-trained segmentation model

- *Segment Anything Model*: Meta, released in 2023: a powerful pre-trained segmentation model that could be used for almost everything
- trained on 11 mil images and their segmentation masks, overs 1 billion object instances
- SAM is not limited to a predefined object classes. You can segment new objects by supplying an example of what we want to do
- the power of SAM comes from the scale of the pre-training dataset (SA-1B)

How SAM works?

- goal of the SA-1B dataset is to create fully segmented images, where each object in an image is given a unique segmentation mask
- each image in the dataset has 100 masks (some of them up to 500 masks - masked objects).
- the process of obtaining the dataset was an increasingly automated data collection
- first, human experts mannualy segmented a small sample of images in order to train an initial model
- the model was used to drive an semi-automated data collection: images first segmented by SAM and next improved by human correction and annotation


NORRMALMSTORG – STRANDV
24
S.S.B

Model functioning

- data on the datasets: (image, prompt, mask)
- **image**: any image
- **prompt**: either a point inside the object to mask or a box around the object to mask
- given the image and the prompt, the model is expected to produce and accurate predicted mask for the object indicated by the prompt, which is compared with the ground truth **mask** label.

Model architecture

- **the image encoder**: similar to the Xception model, takes as input the image and produce a smaller image embedding
- **prompt encoder**: maps prompts to an embedded vector
- **mask decoder**: takes the image embedding and the prompt embedding and outputs a few possible predicted masks. We can compare the predicted masks with the ground truth and get a score
- the components above are trained simultaneously by forming batches of new (image, prompt, mask) triples to train from the SA-1B image and mask data.
- for a given input image, choose a random mask in the input. Next, randomly choose to create either a point or a box prompt. For point prompt, choose a random pixel inside the mask label. for box prompt, draw a box around all points inside the mask label.
- we can repeat the process indefinitely, sampling a number of (image, prompt, mask) triples from each image input.

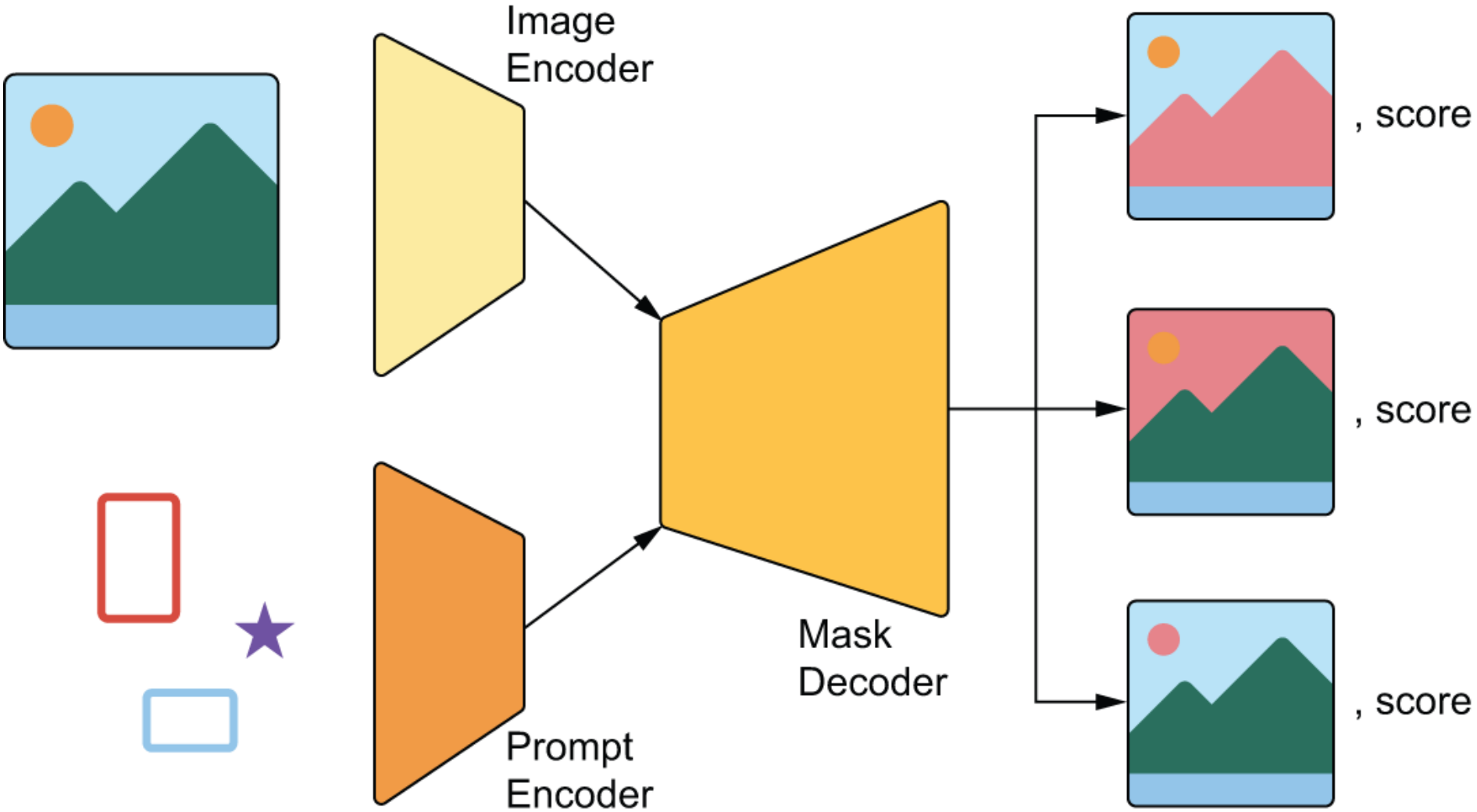


Image
Encoder
Prompt
Encoder
Mask
Decoder
, score
, score
, score