Meta AI

Paper Review

Segment Anything

Author:

Badr Bouzinab

July 2023

Contents

1 Introduction:

Foundation models are large AI models trained on enormous quantities of data in an unsupervised fashion. In NLP, large language models (GPT, BERT...) are pre-trained on web-scale datasets where the pre-training task is to predict the next word in a sentence. This allows the models to learn general language understanding and features from the data. In vision, CLIP uses a text and image encoders to learn meaningful connections between visual and textual representations. With their strong zero-shot and few-shot generalization, foundation models can generalize to new downstream tasks and even perform better than fine-tuned models on that specific task.

This paper introduces a "foundation model" for image segmentation, allowing to solve in zero shot manner a wide range of computer vision tasks such as: edge detection, object proposals and instance detection. To develop this model, 3 questions need to be answered:

1. What pretraining task will enable zero-shot generalization?
2. What is the corresponding model architecture?
3. What data can power this task and model?

2 Segment Anything Task:

The segment Anything task is a promptable segmentation task that:

- Takes as input a prompt specifying what to segment in an image (e.g a rough box or mask, free-form text...).
- Returns a **valid** segmentation mask. "valid" means that even when the prompt is ambiguous and could refer to multiple objects, the output should be a reasonable mask for at least one of those objects.

This task allows for a natural pre-training objective and a general method for zero-shot learning via prompting. In pre-training, given a training sample, we can generate a sequence of prompts refering to the same mask. Then the model compares the predicted mask against the groundtruth, enabling it to be ambiguity-aware. Moreover, downstream tasks can be solved by engineering appropriate prompts. One simple use case is cat instance segmentation. This task can be solved by providing a bounding box detector for cats as a prompt to the model.

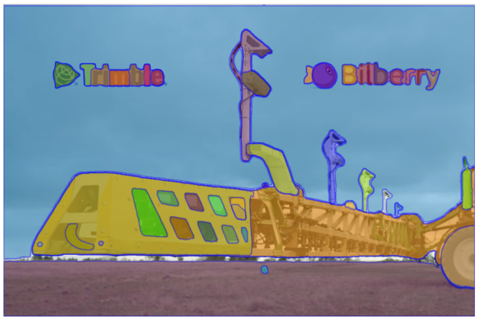

Figure 1: Example image with overlaid masks

3 Segment Anything model (SAM):

As the model is intended for real-world use and user-interactive scenarios, the model architecture must support flexible prompts and be capable of computing segmentation masks in amortized real-time. Therefore, the model is composed of three components:

- Image encoder: This component computes an image embedding. A masked auto encoding (MAE) pre-trained Vision Transformer with slight modifications is used for this purpose.

- Prompt encoder: different encoding strategies are considered depending on the prompt type (an off-the-shelf text encoder from CLIP for text-free prompt, learned positional embedding for points and boxes...)
- Mask decoder: the mask decoder maps the image embedding, prompt embeddings to predict segmentation masks. This decoder is a modified version of a Transformer decoder block. It involves using prompt self-attention and cross-attention in two directions (from prompt to image embedding and vice-versa) to update all embeddings.

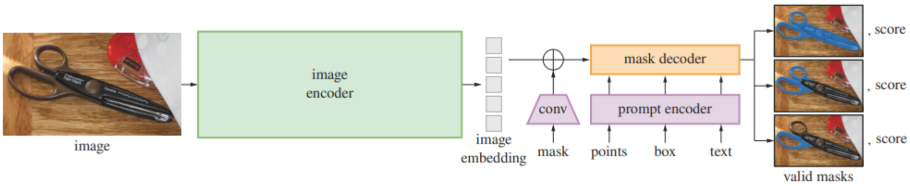

Figure 2: Segment Anything Model (SAM) overview.

4 Segment Anything Data Engine:

As foundation models need to be trained on large amount of data in order to generalize to new data distributions. The Data engine collects a dataset of masks in 3 stages:

In the first stage, SAM assists annotators in annotating masks. In the second stage, SAM automatically generates masks for a subset of objects by prompting it with likely object locations, the remaining objects are annotated by human. In the final stage, a regular grid of foreground points is given as a prompt, and the model outputs on average ∼100 high-quality masks per image. This process yielded on dataset of 11M images and 1B masks.

5 Experiments:

SAM was evaluated on a set of 23 segmentation datasets, producing high-quality masks from a single foreground point prompt, and performing well, often only slightly below that of the manually annotated ground truth. Next, SAM was evaluated on a variety of downstream tasks, including edge detection, object proposal generation, instance segmentation, and text-to-mask prediction, proving the zero-shot transfer capabilities of the model.

6 Conclusion:

In this review, we discussed the Segment Anything paper by Meta. From a personal point of view, I believe that this powerful foundation model for image segmentation, in addition to performing downstream tasks in computer vision, will enable a world of new applications in the near future.

Finally, by providing SA-1B, the largest dataset to date with 400× more masks than any existing segmentation dataset, the SA-1B dataset becomes a valuable resource for the research community to advance the field of computer vision.