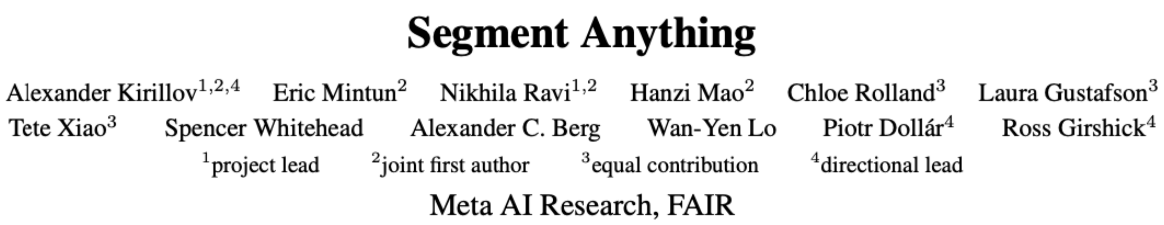


Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

Introduction

The Segment Anything paper has been released in April 2023 by Meta AI and has been quite popular within the CV and ML community. It provides a method for semantic segmentation with high accuracy, and with high zero-shot performances (Segment Anything Model). Part of the paper contribution is the segment anything dataset, containing 11M images annotated with over 1 billion masks.

Paper: https://arxiv.org/pdf/2304.02643.pdfv
Github Repository: https://github.com/facebookresearch/segment-anything
Blog and demo: https://segment-anything.com/

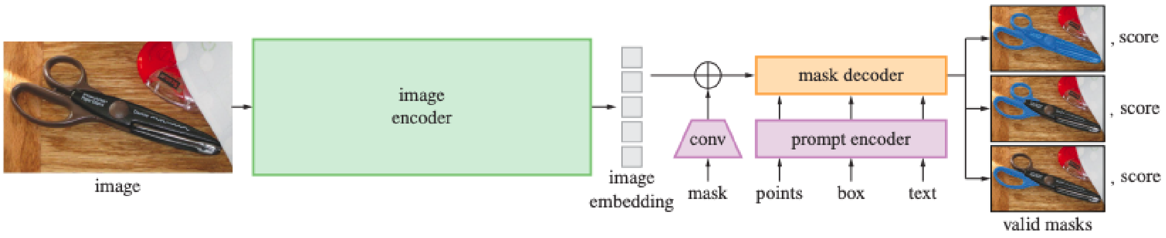
I chose this paper among others, for several reasons:

- The trained model shows an impressive performance, it allows to precisely extract small objects from any type of scene, but the implementation also gives some flexibility by allowing to use prompts to guide the final output. This way, the method can also be guided by prompts deduced from another methods, like object detectors, and hence perform well on related tasks, like instance segmentation, or edge detection.
- The dataset in itself is a major contribution, firstly because it's the larger segmentation dataset to date, but also because the average number of masks per image is notably high. The other interesting aspect of it is the way it was built. The team created an iterative labeling pipeline, made of assisted-manual, semi-automatic, and fully automatic tasks.
- This paper is a good example of the work currently done to use of multimodal foundation models in the field of computer vision. Although those models have been largely democratized in the past few years on language applications, or image generation, there are still many applications in which their potential has not been exploited. It is also a great example of how self-supervised learning can be used to train model reach a high generalization capacity.
- One important aspect of this project is that the data, the scripts and the model weights are open source. This allowed the technology to already be used as a block on other related papers. For example, the inpaint anything paper is based on the SAM segmentation.
- Finally, the model is quite easy to use, and the mask generation is rather fast. When using the interface provided by Meta, the image embeddings are computed once and cached, allowing to fastly recompute masks with different prompts (50ms per computation once the embeddings are computed).

Paper Details

Architecture

The model is composed of an image encoder, based on a masked auto-encoder vision transformer, and a prompt encoder (CLIP is used for the text encoding). The two information are combined and given to a mask decoder that predicts the masks. The mask decoder is a transformer decoder block modified with attention mechanisms adapted to the prompts, followed by a dynamic mask prediction head.



Segment Anything Task

Training is made on promptable segmentation task, where a prompt can be a set of points, a bounding-box, a mask, or a text entry that describes the desired output.
A prompt can lead to ambiguity, for example a single point can simultaneously be part of several objects. To address this, the model is trained to output multiple masks for each prompt, and assign confidence scores based on the estimated IoU to each.

Data engine

To constitute the dataset, a data engine is built, consisting of three stages:

- *Assisted-manual stage*: In this first step, professional annotators click on objects, masks are generated with the assistance of SAM (trained on existing segmentation datasets), and are corrected manually by the annotators. This step helped collecting 4.3M masks on 120k images.
- *Semi-automatic stage*: this step aims at collecting more diverse masks. It starts by providing an automatic detection with the model trained in the previous step, and ask human annotators to annotate objects that were not automatically detected. This step brought 5.9M more masks in 180k images.
- *Fully automatic stage*: The last step has no human intervention. Masks are detected from several point prompts in each image. A selection and a refinement are then done using the multiple masks prediction of the model and the estimated IoU. This final step helped reaching the number of 1.1B masks on 11M images.

Zero-shot transfer experiment

The model is additionally tested on segmentation related tasks, which it was not trained on. Those tasks include edge detection, object proposal generation (segment all objects found), instance segmentation (detect an object and extract a mask), segmentation from free-form text.

Results

- To validate the models performance at mask generation from a single point, IoU is computed and compared to other segmentation models when the ground truth is available. An evaluation of the mask quality is also made by human annotators. Both experiments show that SAM performs better than the state of the art segmentation models on such tasks.

- Regarding edge detection, the observation made is that SAM is able to detect many edges, and even more than the one annotated in the used dataset (BSDS500), lowering the measured precision.
- For the task of object proposal, the model is benchmarked on the LVIS v1 dataset. It is compared to the state of the art method ViTDet-H, which was trained on this dataset. SAM is slightly outperformed by this method but still gives decent results in comparison.
- When performing Instance Segmentation, the same result is observed, SAM is slightly below the baseline's performance, but get comparable accuracy.
- Finally the Text-to-Mask tests show that the model can indeed extract an accurate mask from a text prompt, although it sometimes needs the additional guidance of an input point.

Conclusion

This model shows great performances in promptable segmentation, but can also compete on state of the art models trained on related tasks as edge detection, object detection, etc. Foundation models trained with large datasets, with unsupervised learning methods, have proven to be well performing, and quite versatile, the knowledge learnt can easily be transferrable to related tasks, which make them quite powerful tools.