УДК 004.932.2

OVERVIEW OF THE SAM NEURAL NETWORK MODEL APPLICABLE TO THE PROBLEM OF THE INSTANCE SEGMENTATION

БОБРОВА НАТАЛЬЯ ЛЕОНИДОВНА
к.т.н., доцент
ХАРКЕВИЧ АНТОН ПАВЛОВИЧ,
СТЕЦКО ВАДИМ ЮРЬЕВИЧ
Магистранты
УО «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: Шевалдышева Елена Зигфридовна
к.ф.н., доцент
УО «Белорусский государственный университет информатики и радиоэлектроники»

Аннотация. В данной работе рассмотрена передовая нейросетевая модель SAM (Segment Anything Model), подходящая для осуществления сегментации экземпляров на изображении, изучены ключевые достоинства и недостатки модели, а также произведен обзор наиболее известных модификаций модели.
Ключевые слова. SAM, FAST SAM, Semantic-SAM, YOLO, сегментация экземпляров.

ОБЗОР НЕЙРОСЕТЕВОЙ МОДЕЛИ SAM, ПРИМЕНИМОЙ К ЗАДАЧЕ СЕГМЕНТАЦИИ ЭКЗЕМПЛЯРОВ НА ИЗОБРАЖЕНИИ

Bobrova Natalya Leonidovna,
Kharkevich Anton Pavlovich,
Stetsko Vadim Yurievich

Scientific adviser: Shevaldysheva Elena Zigfrididovna

Abstract. This paper examines the advanced neural network model SAM (Segment Anything Model) suitable for instance segmentation in an image, studies the key advantages and disadvantages of the model, and reviews the most well-known modifications of the model.
Keywords. SAM, FAST SAM, Semantic-SAM, YOLO, instance segmentation.

The task of instance segmentation in an image is one of the key tasks in computer vision. One of the advanced open-source models for instance segmentation task will be discussed in this article. This is the SAM (Segment Anything Model) model. It was released by Meta* AI in 2023 and is very powerful. This is due to a number of its features. The first feature is a very voluminous dataset on which the model was trained. This dataset contains about 11 million images and 1.1 billion object masks. From the first feature of the model follows its second feature. Training on a huge amount of data allowed SAM to acquire the properties of a fundamental model. During training, this model acquires the ability to solve many other tasks, for which it was not initially trained. Thus, SAM is able to independently segment a set of objects in the image, as well as produce the

segmentation result at the user's request. The user can mark a point or several points of the object, mark the bounding box of the object, color the approximate area of the object and even write a text description of the object. In addition, the model can be easily pre-trained for a specific task or subject area.

(*Recognized as an extremist organization in the Russian Federation).

Let's consider the model architecture, which is presented in figure 1:

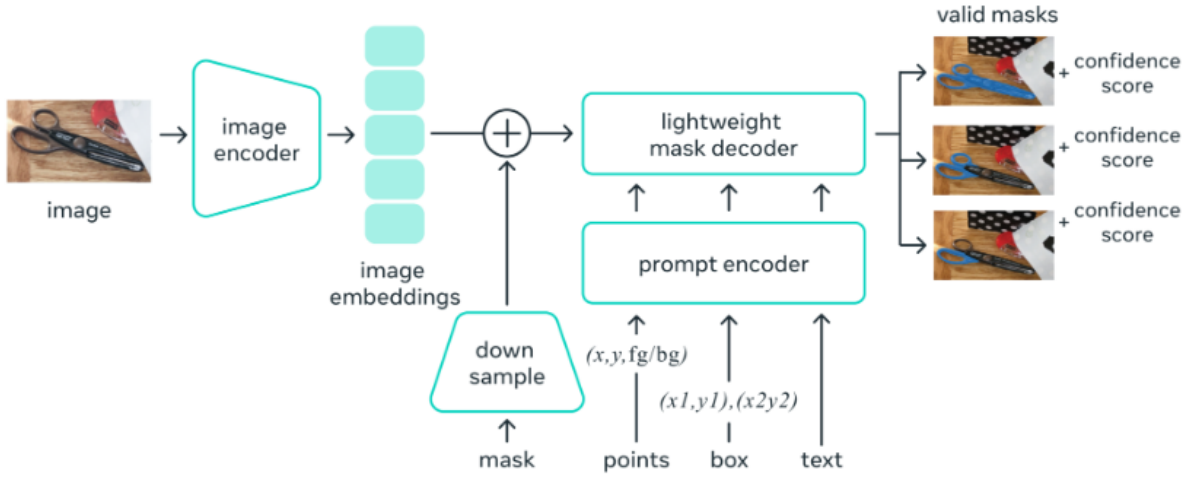

Fig. 1. The architecture of SAM

The input image is input of the image encoder, which returns an embedding. Embedding is a low-dimensional representation of the image. Next comes the encoder of additional information specified by the user: object mask, point, bounding box, or text description of the object. The last component of the model is the decoder, which uses the image embeddings and additional user information as input. The peculiarity of the model is that it outputs not one segmentation mask, but three. This is necessary to handle boundary cases, for example, when a point of a marked object can refer to several objects in the image at once. The number three was derived empirically by the developers.

The mechanism of generating three segmentation masks for one point is shown in Figure 2:

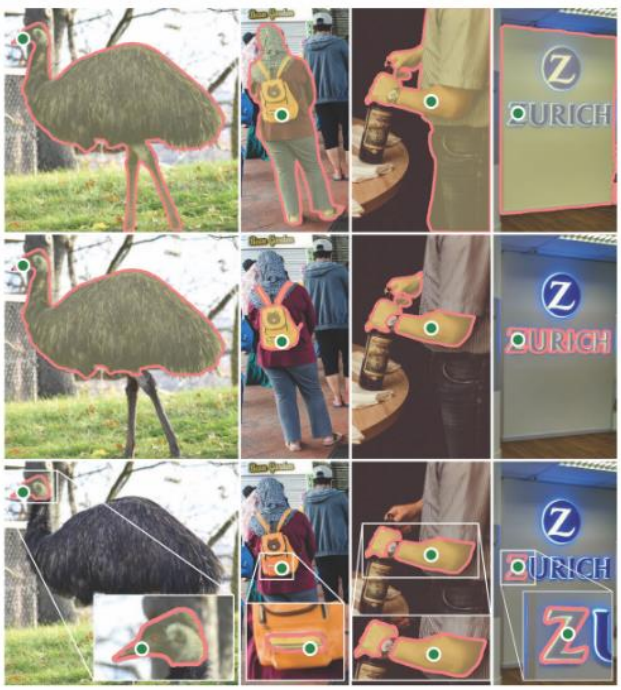

Fig. 2. The mechanism of generating three segmentation masks for one point

The SAM model already has several modifications that are worth paying attention to. When comparing the SAM model with the YOLOv8n-seg model, SAM is much inferior to YOLO in size and speed (358 MB for SAM vs. 6.7 MB for YOLO, 51096 ms/im for SAM vs. 59 ms/im for YOLO). The idea of incorporating the YOLOv8n-seg architecture into the SAM architecture led to the creation of the FAST SAM model.

FAST SAM differs from SAM in that it is much faster without losing segmentation quality. This was achieved by replacing the SAM encoder with a CNN detector. All objects in the image are segmented and selected according to the user's request in the interactive mode of the model. The segmentation of all objects is based on YOLOv8n-seg architecture, which is the reason for the speed improvement. When comparing FAST SAM with YOLOv8n-seg we have comparable size and speed of the models (23.7 MB for FAST SAM vs. 6.7 MB for YOLO, 115 ms/im for FAST SAM vs. 59 ms/im for YOLO).

The architecture of the FAST SAM model is shown in Figure 3:

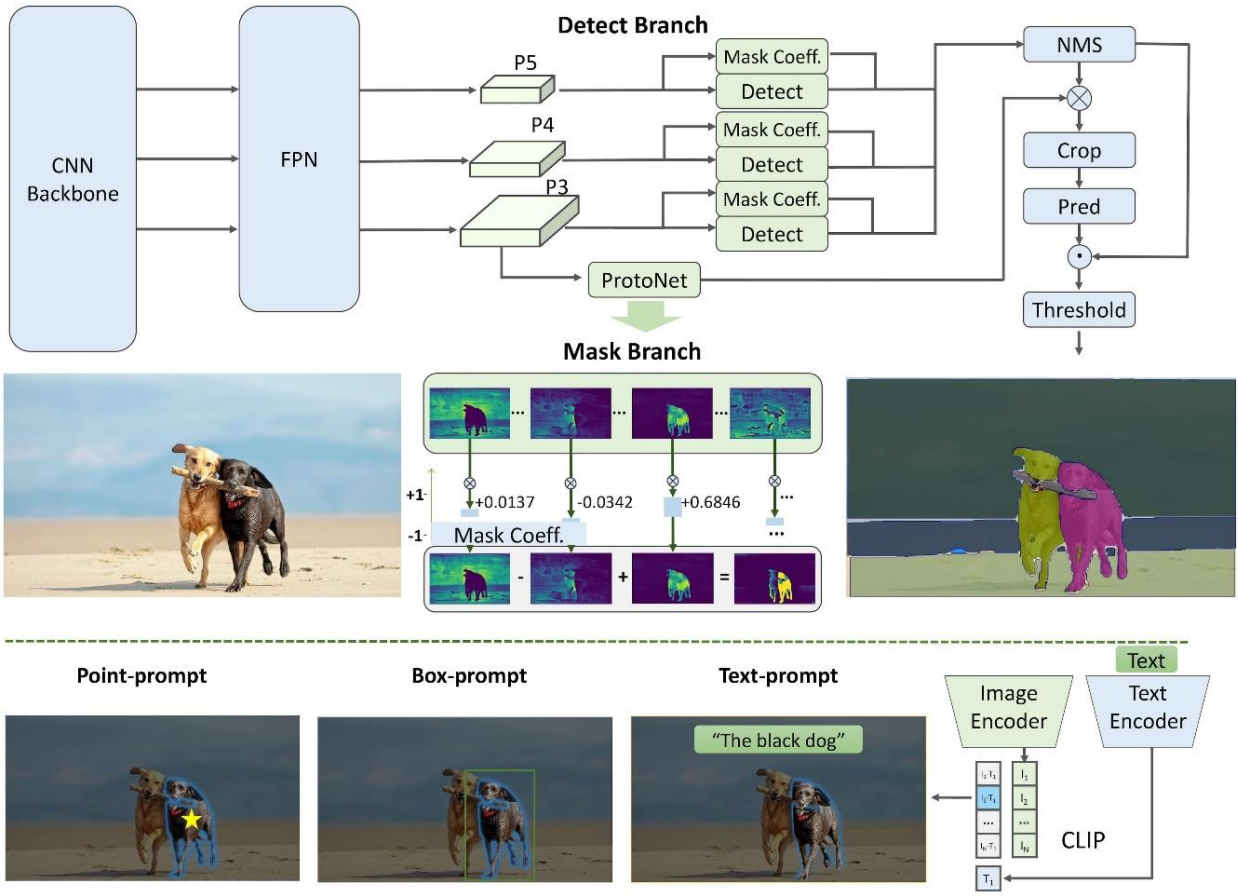

Fig. 3. The architecture of FAST SAM

FASTER SAM was introduced in the context of the idea of running SAM on low-performance devices. The encoder was changed like in the FAST SAM model. It was replaced by a lighter weight ViT-S encoder. A distillation process was used between the original encoder and the lightweight decoder in order to maintain the necessary coupling between the encoder and the mask decoder. Distillation is a technique used to train a new model by transferring knowledge from another pre-trained model.

The architecture of the FASTER SAM model is shown in Figure 4.

The Semantic-SAM model was created to extend the level of segmentation granularity. It allows model to segment object at all levels of granularity. The model was trained on 6 datasets in addition to the original one in order to achieve such results. Different datasets contain segmentation masks at different levels of granularity. Also, the predicted object segmentation masks in the model training algorithm are compared not with a single true object segmentation mask, but with several masks at different levels of detail.

The architecture of the Semantic-SAM model is presented in Figure 5.

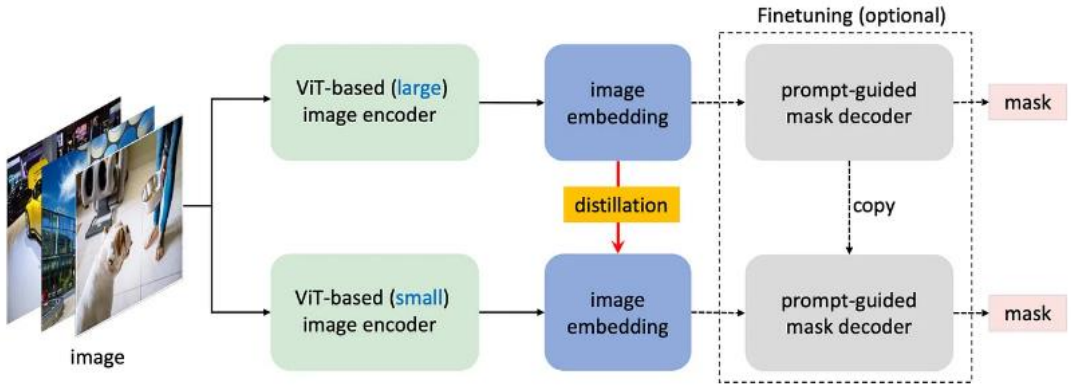

Fig. 4. The architecture of FASTER SAM

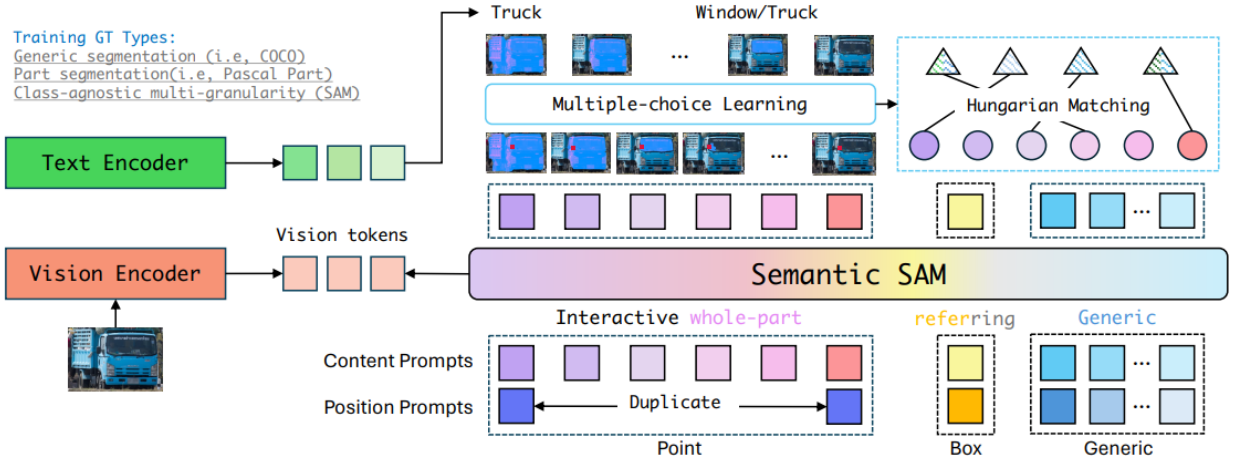

Fig. 5. The architecture of Semantic-SAM

In conclusion, it is worth noting the following. Despite the fact that SAM has some disadvantages (need for pre-training for a specific task, low speed compared to models such as YOLO), it is an advanced model in the field of segmentation. It has fundamental qualities, and its modifications allow developers to expand the range of model applications depending on the requirements to the level of detail and speed of operation, which makes SAM an even more versatile model.

References

1. Chaoning Zhang, Yu Qiao, Shehbaz Tariq, Sheng Zheng, Chenshuang Zhang, Chenghao Li, Hyundong Shin, Choong Seon Hong “Understanding Segment Anything Model: SAM is Biased Towards Texture Rather than Shape”// arXiv.org, 3 June 2023

2. Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, Jinqiao Wang “Fast Segment Anything”// arXiv.org, 21 June 2023

3. Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, Choong Seon Hong “Faster Segment Anything: Towards Lightweight SAM for Mobile Applications”// arXiv.org, 25 June 2023

4. Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, Jianfeng Gao “Semantic-SAM: Segment and Recognize Anything at Any Granularity”// arXiv.org, 10 July 2023