UDC 004.932.2

OVERVIEW OF THE SAM 2 NEURAL NETWORK MODEL APPLICABLE TO THE PROBLEM OF THE INSTANCE SEGMENTATION IN VIDEO

BOBROVA NATALYA LEONIDOVNA,
candidate of technical sciences, associate professor
KHARKEVICH ANTON PAVLOVICH,
STETSKO VADIM YURIEVICH
undergraduates
Belarusian State University of Informatics and Radioelectronics

Scientific adviser: Shevaldysheva Elena Zigfrididovna
ph.d., associate professor
Belarusian State University of Informatics and Radioelectronics

Аннотация. В данной работе рассмотрена нейросетевая модель SAM 2(Segment Anything Model 2), подходящая для осуществления сегментации экземпляров на видео, произведено сравнение SAM 2 с предшествующей моделью SAM, рассмотрена архитектура модели SAM 2, её ключевые компоненты, процесс обучения модели.
Ключевые слова. SAM 2, SAM, сегментация экземпляров, маски сегментации, сегментация объектов в реальном времени, последовательность кадров.

ОБЗОР НЕЙРОСЕТЕВОЙ МОДЕЛИ SAM 2, ПРИМеНИМОЙ к ЗАДАЧЕ СЕГМЕНТАЦИИ ЭКЗЕМПЛЯРОВ НА ВиДЕО

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»

Abstract. This paper reviews the neural network model SAM 2(Segment Anything Model 2) suitable for instance segmentation in a video, compares SAM 2 with the previous SAM model, reviews the architecture of SAM 2 model, its key components and the training process.
Keywords. SAM 2, SAM, instance segmentation, segmentation masks, real-time object segmentation, frame sequence.

The SAM 2 model considered in this paper is the successor of the SAM model. It is a unified model developed for real-time object segmentation on both images and video [1].

The key differences between the SAM 2 model and the SAM model are:

1) The SAM 2 model dataset has grown 4.5 times compared to the past model.
2) The number of annotation for video segmentation has increased by 53 times.
3) SAM 2 segments the image 3 times faster than SAM with superior accuracy.
4) SAM 2 has learnt how to handle complex scenarios where objects move or change appearance quickly.

It has been found in research papers that SAM 2 is significantly more accurate compared to its predecessor model [2].

An example of the different instance segmentation accuracy of SAM and SAM 2 models is shown in Figure 1 below:

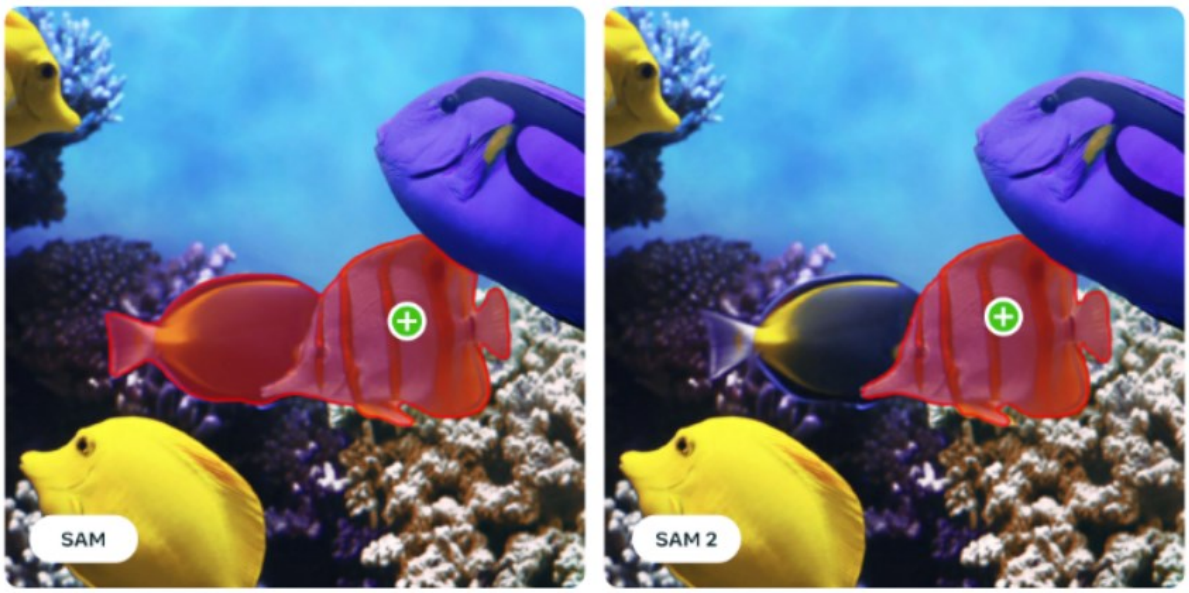

Fig.1 Different instance segmentation accuracy of SAM and SAM 2 models for the image with a point hint

SAM 2 architecturally extended the SAM model to work with both video and images. During video segmentation, SAM 2 uses point or mask hints on a single frame, then determines the spatial extent of an object and segments it throughout the video. The model works similarly to SAM for image processing. The lightweight mask decoder uses any hints to generate a segmentation mask.

The SAM 2 architecture is shown in Figure 2 below:

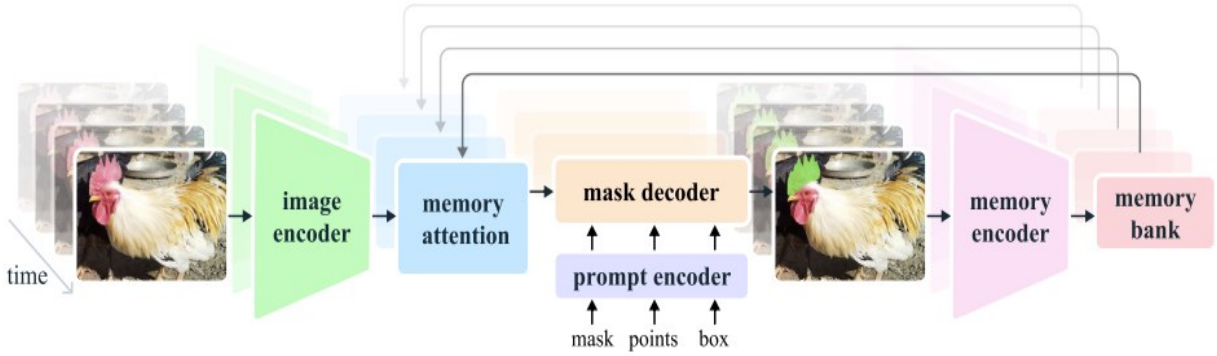

Fig.2 The architecture of the SAM 2 model

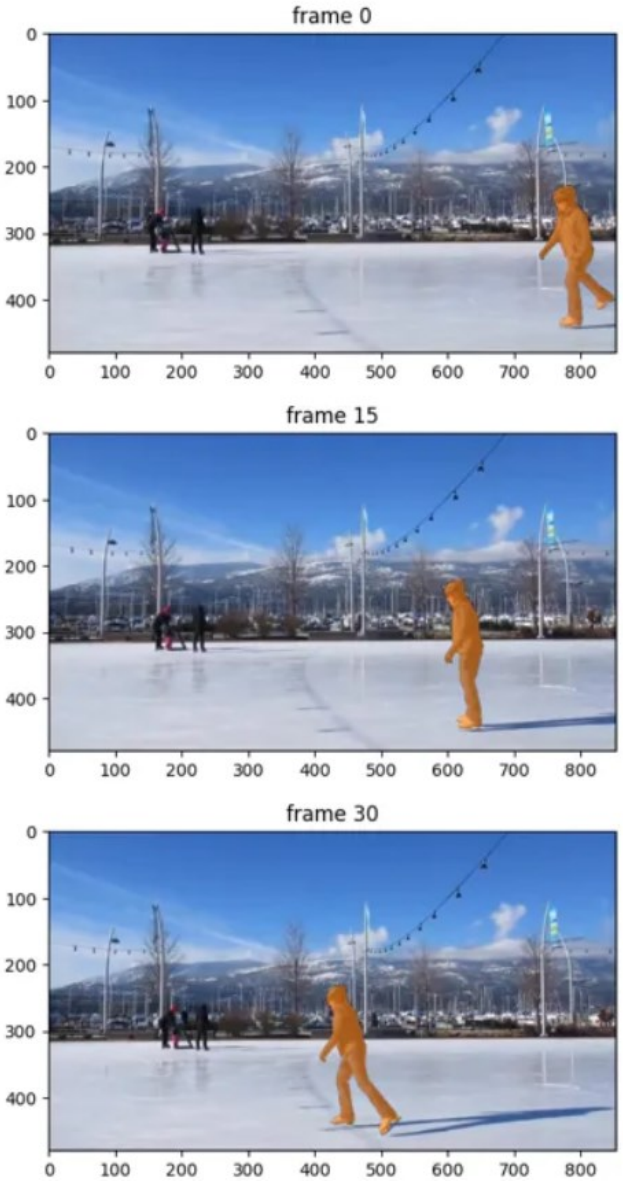

Fig.3 An object segmentation in motion on a sequence of frames using the SAM 2 model

The figure above shows that the segmentation prediction in each frame is based on the current hint and any previously observed similar frames.

The videos are processed in streaming mode, with frames being analysed one at a time by the image encoder, which makes reference to memories of the target object from earlier frames. The image encoder also processes each video frame to create feature embeddings, which are essentially compressed representations of the visual information in each frame. The image encoder is run only once for the entire video, making it very efficient.

The memory attention block helps the model use information from previous frames and any new hints to improve segmentation of the current frame. It uses a number of transducer blocks to process features of the current frame, compares these features with memories of past frames, and updates the segmentation mask based on both new features and previous memories. In this way, the memory attention block helps to handle complex scenarios in which objects may move or change over time.

Like the hint encoder in the SAM model, a hint encoder in SAM 2 accepts any input hints to determine which part of the frame to segment. It uses these hints to refine the segmentation result.

The mask decoder, which can also use input hints, predicts a segmentation mask for the frame. If the hint is unclear, it predicts several possible masks and selects the best one based on overlap with the object.

The memory encoder transforms the prediction and embeddings of the image encoder for use in future frames. It creates memories of past frames by summarising and combining information from previous masks and the current frame. This helps the model to remember and use information from earlier videos.

The memory bank stores memories of past frames and hints. This includes a queue of recent frames and hints, as well as high-level information about objects. It helps the model track changes and movements of objects over time.

An example of an object segmentation in motion on a sequence of frames using the SAM 2 model is shown below in Figure 3.

The training of the SAM 2 model proceeded as follows: the model learnt to predict segmentation masks by interacting with sequences of frames. It received various hints (point, borders, masks) to manage its predictions. This helped the model to respond better to different types of input data and improve segmentation accuracy. Object segmentation accuracy was improved based on iterative hints. The training of SAM 2 was broken down into 3 phases:

1) In the first phase, human annotators and the SAM model were used for annotation. SAM first generated object masks on the frames, then humans corrected and refined the annotations with pixel accuracy. The video frames used for annotation were extracted at a rate of 6 frames per second. SAM 2 was then trained on the annotated data.

2) The second stage also used human annotators and the SAM model. In addition, the SAM 2 model trained in the first stage was used in this stage. First, humans and SAM marked up the frames in a similar manner to the first stage. Then, the SAM 2 model generated object masks for the entire video by temporally propagating hints from the first frame. Spatio-temporal masks were generated for the entire video in this way. With the SAM 2 model, the annotation process in the second stage took 5 times less time than in the first stage. At the end of the second stage, the SAM 2 model was further trained using 63,500 masks.

3) In the last phase, only the SAM 2 model was used to generate hints. Human intervention was minimal in specific cases. 197,000 masks were generated in the last phase.

A separate group of annotators performed mask validation at each step. They checked whether the masks themselves were selected correctly and how well the mask was selected over the length of the whole video. Anything that was poorly selected was re-mapped, anything that was highly ambiguous was discarded.

Thus, the SAM 2 model is fundamental and has great potential for application. Its ability to segment both static and dynamic visual data makes it a versatile tool for researchers and developers. It is already starting to be used for segmentation of medical images [3,4]. The model can also be used in augmented and virtual reality applications. For example, SAM 2 can accurately identify and segment real-world objects and make interaction with virtual objects more realistic.

References

1. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, Christoph Feichtenhofer “SAM 2: Segment Anything in Images and Videos”// arXiv.org, 28 October 2024

2. Osher Rafaeli, Tal Svoray, Roni Blushtein-Livnon, Ariel Nahlieli “Prompt-Based Segmentation at Multiple Resolutions and Lighting Conditions using Segment Anything Model 2”// arXiv.org, 15 August 2024

3. Jiayuan Zhu, Yunli Qi, Junde Wu “Medical SAM 2: Segment medical images as video via Segment Anything Model 2”// arXiv.org, 1 August 2024

4. Haoyu Dong, Hanxue Gu, Yaqian Chen, Jichen Yang, Yuwen Chen, Maciej A. Mazurowski “Segment anything model 2: an application to 2D and 3D medical images”// arXiv.org, 22 August 2024