Segment Anything Model Fine-tuning on Electron Microscopic Images of Cells' Mitochondria

Authors:

Anh Thu DOAN
Amr MOHAMED
Gwendal AUPHAN

AI BASED IMAGE PROCESSING MODULE

ENGINEERING SCHOOL'S THIRD YEAR, AI SPECIALIZATION, GROUP 2

Monday 6th November, 2023

Abstract

In this paper, we go through the process of fine-tuning Meta AI's Segment Anything state-of-art model in segmentation tasks on EPFL's Electron Microscopy Dataset, that contain cellular level images contacting mitochondria and the respective segmentation masks, with the aim to direct the SAM model to excel at the task of segmenting the mitochondria in the cellular images. We, therefore, go through the process of data preparation, fine-tuning plan conception, loss functions and evaluation metrics choice, and finally, model evaluation on a given test set.

Contents

1 Introduction

In the rapidly evolving landscape of artificial intelligence, breakthroughs continue to redefine the boundaries of what is possible. Meta AI's "Segment Anything Model" (SAM) [1] represents a cutting-edge innovation that promises to revolutionize image processing by enabling the precise segmentation of objects in images. While SAM holds immense potential across various domains, this report focuses on its application to the realm of medical imaging, specifically its use in the segmentation of mitochondria.

The motivation behind this study stems from the inherent challenges presented by medical image analysis, particularly in the domain of mitochondrial research. Mitochondria, often called the "powerhouses of the cell," play a pivotal role in cellular function, energy production, and even apoptosis regulation. Studying the morphology and dynamics of mitochondria is crucial for understanding cellular health, metabolism, and various diseases, including neurodegenerative disorders, cancer, and metabolic disorders.

Traditionally, the manual annotation of mitochondria in electron microscopy (EM) images has been a labour-intensive and time-consuming task, hampering the research progress in this field. Automated segmentation methods have the potential to greatly expedite this process, making it more efficient and less prone to human error. This is where SAM enters the picture.

However, SAM, in its out-of-the-box state, falls short of delivering the desired accuracy when applied to medical images, particularly those containing intricate structures like mitochondria. The inherent complexity of mitochondria's morphology, their diverse sizes, shapes, and textures, and the often noisy and low-contrast nature of EM images pose substantial challenges for SAM. As a result, fine-tuning this model becomes imperative to adapt it to the specific demands of this domain.

This report aims to detail the efforts undertaken to enhance the performance of SAM on mitochondria segmentation. We explore the key motivations behind this endeavour, emphasizing the significance of precise mitochondrial segmentation for medical research and how SAM's capabilities can potentially address the ex-

isting limitations. The subsequent sections will delve into the methodology employed, the dataset used for fine-tuning, the experimental setup, and the results obtained.

2 Methods

In this section, we go through the several project steps of defining the dataset used, the data preparation steps, the model's architecture, fine-tuning plan, loss functions, and the evaluation metrics used to evaluate the model's performance on the test set, and comparing it with the base SAM model (before the fine-tuning).

2.1 Dataset Description

In this study, we use ***EPFL's Electron Microscopy Dataset***, containing images which comprise 5x5x5µm sections extracted from the CA1 hippocampus region of the brain, with the respective segmentation masks of the mitochondria. The choice of this specific dataset is motivated by the significance of mitochondria within neural tissues and the need for precise segmentation within a biologically relevant context.

2.2 Data Preparation

Before the fine-tuning process, several essential data preparation steps were undertaken to optimize the model's performance:

- Dividing Images into Patches[5]
  - Number of Training Images: To increase the datasetś diversity and to address the inherent challenges of mitochondrial segmentation, a set of 165 original images, each with dimensions of (768, 1024) pixels, were divided into smaller patches.
  - Image Patches: This process yielded 1980 training images, each with dimensions reduced to (256, 256) pixels. Generating these smaller patches facilitates the model's ability to capture fine-grained details and increases the amount of training data.
- Rescaling

- Image Intensity Rescaling: To ensure consistency in the model's inputs, all images were rescaled to a pixel intensity range between 0 and 255. This standardization allows the model to effectively learn features across the entire dataset without being biased by variations in pixel intensity.
- Mask Rescaling: Additionally, binary masks representing the target segmentation labels were rescaled to a binary range of 0 and 1. This transformation aligns the mask values with the segmentation task, with 0 representing the background and 1 signifying the mitochondria of interest.

2.3 SAM Model

2.3.1 Definition of SAM Model

In the realm of Computer Vision (CV), the introduction of Meta AI's Segment Anything Model (SAM) has ignited a profound sense of enthusiasm and intrigue. SAM is an image segmentation model with a remarkable capacity to produce segmentation masks, responding adeptly to diverse input prompts. One of SAM's most remarkable attributes is its unparalleled zero-shot transfer capabilities, showcasing its proficiency across an extensive spectrum of tasks and datasets. Undoubtedly, SAM's exceptional achievements mark a significant milestone in the evolution of foundational models within the field of Computer Vision, underscoring its pivotal role in shaping the future of CV.

- Training Data: SAM was trained on a vast dataset of over 11 million images, with more than 1 billion corresponding masks. This extensive training data equips SAM with a broad understanding of object segmentation.
- Versatile Input: SAM is designed to accept various human prompts for object segmentation. It can effectively process prompts through points, bounding boxes, or even text descriptions, allowing for flexible and user-friendly interaction with the model. This adaptability makes it a powerful tool for diverse segmentation tasks.

2.3.2 SAM's features

- Zero-Shot Generalization: SAM's capability to generalize and segment objects it has never encountered before, without additional training, expands the horizons of segmentation tasks. This attribute is particularly valuable when dealing with diverse and complex objects like mitochondria.

- Flexible Prompting: SAM is designed to accept various input types, including points, bounding boxes, and text descriptions. This flexibility allows researchers to interact with the model in a way that best suits the specific segmentation task at hand.

- Real-Time Mask Computation: The ability of SAM to generate masks for objects in real-time positions is an ideal tool for applications where quick and accurate object segmentation is crucial, such as in autonomous driving and robotics.

- Ambiguity Awareness: SAM's awareness of object ambiguity in images is a significant advantage. It can generate masks for objects even when partially occluded or overlapping with other objects, enhancing its suitability for challenging segmentation scenarios.

2.3.3 SAM architecture

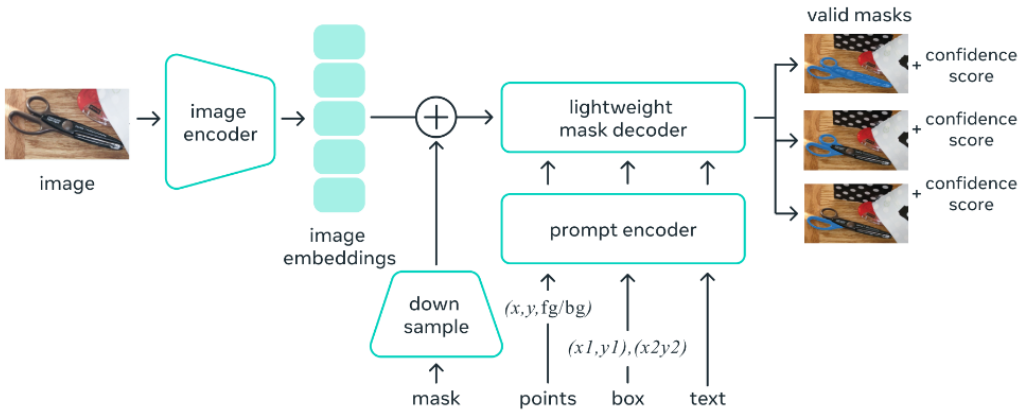

Figure 1: Segment anything model Architecture.

- Image Encoder: SAM initiates the segmentation process by encoding the input image into a high-dimensional vector representation. This encoding is performed using a Vision Transformer (ViT-H) model, which has undergone pre-training on an extensive image dataset. The high-dimensional vector representation captures the salient features of the input image.
- Prompt Encoder: Simultaneously, SAM encodes the user's prompt into a separate vector representation. The prompt can be provided in various forms, such as points, bounding boxes, or text descriptions, and this encoding enables SAM to understand the user's intent accurately.
- Mask Decoder: These two vector representations, representing the image and the prompt, are combined and directed to a mask decoder. The mask decoder, a lightweight transformer model, is responsible for generating a binary mask that corresponds to the object specified in the user's prompt. This mask identifies the segmented object within the input image.

2.3.4 SAM model fine tuning

To fine-tune the SAM model, we then define the fine-tuning plan. As represented in the previous section, The image encoder and prompt encoder are two different encoders responsible for taking the vector representation of the images or the prompts and mapping them to higher dimensional vector spaces. Since these two building blocks can be thought of as prior information retrieval phases, and they already have been trained for a long time by Meta AI on various domains, fine-tuning them was found not to be helpful, and we can risk the loss of important weights of the encoders by fine-tuning them. Therefore, prior to the fine-tuning phase, we **freeze the image and prompt encoders** so that we force no updates of the weights to be performed. However, since the mask decoder is the building block that's responsible for combining the output of the two encoders and building the relationship rules between a given cellular image and the prompt (bounding box(s)) surrounding the mitochondria(s), therefore we **allow the weights update of the mask decoder**.

The fine-tuning of the model was found to be computation resources exhaustive, and by applying several training plans, it was found that **five epochs** was a good number of fine-tuning epochs to prevent overfitting on the cellular data by longer times training phases, and also to prevent resources consumption with no significant model improvement.

2.4 Optimizer and Loss function

2.4.1 Optimizer

The optimizer plays a central role in shaping how our model learns and adapts. In this context, we leverage the versatile and widely recognized Adam optimizer.

The choice of the Adam optimizer is strategic. Adam combines the advantages of two optimization techniques, AdaGrad and RMSprop, making it well-suited for various training scenarios. It offers dynamic learning rate adjustments, which help the model converge more efficiently during training. The learning rate, set at 1e-5, ensures a balanced rate of parameter updates, preventing overshooting and

instability during the optimization process.

Moreover, by setting weight decay to zero, we emphasize the importance of fine-tuning the model parameters without introducing regularization. This configuration aligns with our objective of tailoring the training process to maximize the model's performance.

2.4.2 Loss function

We adopt a custom loss function from the **monai package**, which combines the Dice loss and Cross-Entropy Loss. This custom loss function simultaneously computes the Dice loss and the Cross-Entropy Loss and returns their weighted sum. This approach offers several advantages:

- Pixel-wise Classification: The Cross-Entropy Loss component ensures that the model becomes proficient in determining the content of each pixel, a vital aspect in distinguishing objects from the background. It plays a critical role in scenarios with a substantial imbalance between background and object pixels.
- Object Boundary Handling: This combined loss function balances accurate pixel classification and precise object boundary delineation. It strengthens the model's ability to make accurate predictions and effectively identify objects in the image.
- Robustness and Versatility: This approach bolsters the model's overall robustness, reducing the likelihood of errors and enhancing its adaptability to diverse datasets during training. It contributes to improved model performance when dealing with different data sources and varying segmentation challenges.

2.5 Model Evaluation

Since the project aims to direct SAM's power from performing segmentation tasks across various domains into our domain of interest, Cellular Segmentation, we

need solid quantification metrics of the model's performance in order to measure the difference in performance between the model before the fine-tuning and after. Therefore, we here define the two different metrics used:

- Sørensen–Dice coefficient[2][3]: It is a statistic used to measure the similarity or overlap between two sets. This is applied in our case to find the proportion of intersection between two segmentation masks, the ground truth mask and the predicted mask. Its formula is given by:

$$\frac{2 \cdot |A \cap B|}{|A|+|B|}$$

  Where:

  - A represents the ground truth mask.
  - B represents the predicted mask.
  - $|A \cap B|$ denotes the size of the intersection of masks in pixels of A and B.
  - $|A|$ is the size of the ground truth mask A.
  - $|B|$ is the size of the predicted mask B.

  The metric provides a measure of how well two sets overlap or agree. It quantifies the proportion of overlap between the sets relative to their sizes. A Dice coefficient of 0 indicates no overlap, while a value of 1 indicates a perfect match or complete overlap. It is also sensitive to small overlaps and can penalize partial overlaps, making it a suitable metric for tasks where fine-grained segmentation is important.

- IoU (Intersection over Union) Score[4]: The IoU score is a metric used to assess the overlap or similarity between two sets, typically applied to evaluate the intersection between two segmentation masks: the ground truth mask and the predicted mask. Its formula is given by:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

  Where:

- A represents the ground truth mask.
- B represents the predicted mask.
- $|A \cap B|$ denotes the size of the intersection of masks in pixels between A and B.
- $|A \cup B|$ represents the size of the union of masks in pixels between A and B.

The IoU score also measures the overlap between two sets, specifically the intersection relative to the union of the sets. It quantifies the proportion of overlap while considering the size of the combined regions. An IoU score of 0 indicates no overlap, and a score of 1 represents a perfect match or complete overlap. It is often considered a balanced metric as it gives equal weight to both false positives and false negatives, making it well-suited for evaluating object detection and localization tasks.

3 Results

In this section, we investigate the results of both the instances of the SAM model, the base (pre-trained) model with the weights from Meta AI, and the fine-tuned model on the cellular images to understand and quantify the performance change in the model's segmentation predictive capacity.

By testing both instances of the model on the test set, the fine-tuned model showed a huge advancement over the pre-trained base model in the model performance after the fine-tuning, where the dice score increased by approximately 0.3 from the untuned to the tuned model, with a total of approximately 0.85, meaning that on average, approximately 85% of the pixels in the test masks overlap with the respective predicted masks by the fine-tuned SAM. In contrast, the IoU score passed from 0.47 on the untuned model to 0.764, indicating that approximately 76.4% of the region where the two masks (the actual and the predicted) overlap corresponds to the same object or segmentation area.

Noticing this huge advancement from the base model (untuned) to the tuned model, we here inspect some examples of the test set to infer and have a better interpretation of the results.

From the figure below, we can see the big difference in performance between the two instances of the model on a randomly selected set of testing images and masks, where the fine-tuned model achieves much higher performance than the base model, more noticeable on the images that have more than one mitochondria. Moreover, in images with only one mitochondrion to segment, the base model already showed a good result. However, there was always at least a slight difference in performance in favour of the fine-tuned instance of the model.

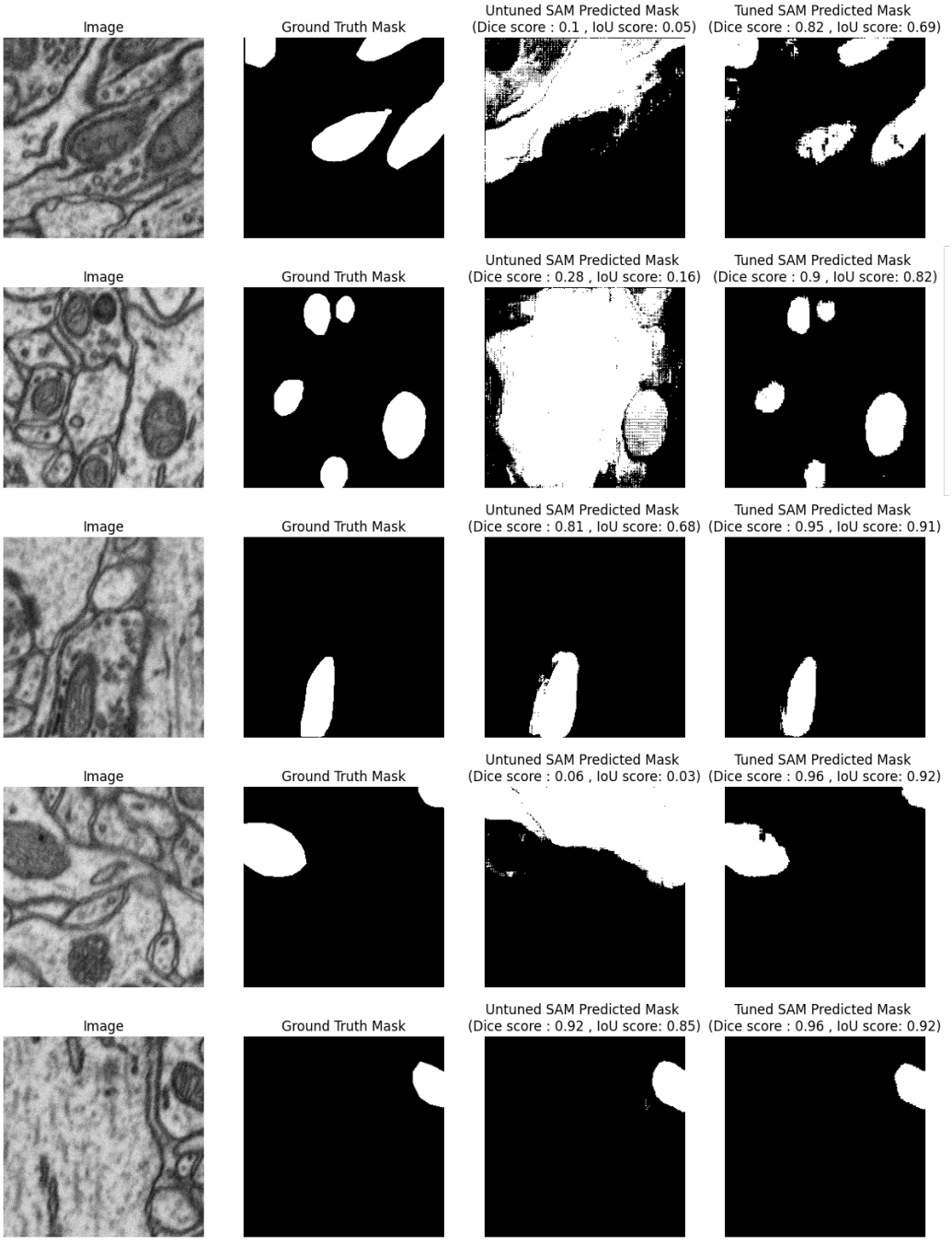

Figure 2: Comparison of the segmentation results of the fine-tuned and base model. Left Panel: Original Cellular Image; Second Panel: Ground Truth Mask; Third Panel: Base Model Prediction; Fine-Tuned Model Prediction

4 Discussion

From the previous section, we can clearly notice the significant improvement in the segmentation predictive capacity of the SAM model after fine-tuning. The two instances of the SAM model, the pre-trained base model with Meta AI's weights and the fine-tuned model on cellular images, were rigorously evaluated on a test set to quantify the performance change.

The most notable finding is the substantial increase in the dice score from approximately 0.55 in the untuned model to around 0.85 in the tuned model. This remarkable improvement suggests that the fine-tuning process greatly enhances the model's ability to segment cellular images accurately. The dice score signifies that, on average, approximately 85% of the pixels in the test masks overlap with the predicted masks by the fine-tuned SAM, signifying that, on average, approximately 85% of the pixels in the predicted masks produced by the fine-tuned SAM model closely match the true positions of mitochondria in the cellular images. In other words, the fine-tuned model is much more adept at accurately segmenting mitochondria, capturing a larger portion of these structures with precision.
In addition, the IoU score also experienced a significant boost, increasing from 0.47 in the untuned model to 0.764 in the fine-tuned model, where the score of 0.47 in the untuned model implies that only 47% of the region where the predicted and actual masks overlap corresponds to the same object or segmentation area. This means that the untuned model struggled to precisely capture the boundaries and shapes of the mitochondria in the images, resulting in a relatively low degree of overlap with the ground truth, while the significant improvement to an IoU score of 0.764 in the fine-tuned model is highly significant, and suggests that approximately 76.4% of the regions where the predicted masks and the true masks overlap now correspond to the same object or segmentation area. In other words, the fine-tuned model aligns its predicted masks much more closely with the actual positions and shapes of the mitochondria in the cellular images.

Furthermore, to gain a deeper understanding of the results, we examined specific examples from the test set. The visual comparison of randomly selected testing images and masks provided valuable insights, where it was evident that the fine-tuned model consistently outperformed the base model, especially in images with

multiple mitochondria to segment. While the base model already showed good results in images with a single mitochondrion, the fine-tuned instance consistently exhibited a slight but noticeable performance advantage.

Implications for Cellular Image Analysis: The observed performance improvement in the fine-tuned SAM model has important implications for cellular image analysis. Accurate segmentation of cellular structures is critical in various scientific and medical applications, including cell biology and disease diagnosis. The ability to automatically and accurately identify and segment mitochondria can significantly reduce the manual interaction required in such tasks, where these tasks are usually performed by domain specialists and experts, and using such a powerful tool introduced in this paper can save a lot of time on the specialists, where their work can be more of a validation work of the images than an annotation work.

Limitations and Future Work

The main limitations of the study were mostly related to the computation resources since fine-tuning SAM efficiently takes long times ranging between 5 to 15 minutes for one epoch on the training set used in the study, limiting the option of exploring different architectures, parameters, loss functions, and optimizers. Therefore, future work can be addressed to investigate further the effect of the variability in each of the components mentioned in the model's predictive power.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick. (2023). "Segment Anything." arXiv preprint, arXiv:2304.02643 [cs.CV].
https://arxiv.org/abs/2304.02643

[2] Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons." Kongelige Danske Videnskabernes Selskab, 5(4), 1–34.

[3] Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species." Ecology, 26(3), 297–302. doi:10.2307/1932409. JSTOR 1932409. S2CID 53335638.

[4] Jaccard index. (2023, September 16). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Jaccard_index

[5] Group and Crowd Behavior for Computer Vision. (2020). Shuyu Sun and Tao Zhang. ISBN 978-0-12-820957-8. Gulf Professional Publishing. Retrieved from https://doi.org/10.1016/C2019-0-02019-5