Lung Segmentation in Chest X-ray Images using the *Segment Anything Model* (SAM)

Gabriel Bellon de Carvalho
Department of Computing
Federal University of São Carlos – UFSCar
18052-780, Sorocaba, SP – Brazil
Email: gabrielbellon@estudante.ufscar.br

Jurandy Almeida
Department of Computing
Federal University of São Carlos – UFSCar
18052-780, Sorocaba, SP – Brazil
Email: jurandy.almeida@ufscar.br

***Abstract*—The *Segment Anything Model* (SAM), introduced by Meta AI in April 2023, represents a cutting-edge tool designed to identify and separate individual objects within images through semantic interpretation. The advanced capabilities of SAM stem from its training on millions of images and masks. Shortly after its release, researchers began evaluating the model's performance on medical images. With a focus on optimizing work in the healthcare field, this study proposes using SAM to evaluate and analyze X-ray images. To enhance the model's performance on medical images, a transfer learning approach was employed, specifically through fine-tuning. This adjustment led to a substantial improvement in the evaluation metrics used to assess SAM's performance compared to the masks provided by the datasets. The results achieved by the model after fine-tuning were satisfactory, demonstrating performance close to that of renowned neural networks for this task, such as U-Net.**

I. Introduction

Segment Anything Model (SAM) [1] is a tool that, since its release, has proven to be very promising in the task of image segmentation. Its approach involves using a variety of input prompts to identify different objects in images, such as points and bounding boxes. To predict the masks, SAM uses three components: (*i*) image encoder, (*ii*) prompt encoder, and (*iii*) mask decoder. Additionally, the model can automatically segment anything in an image and generate multiple valid masks for ambiguous inputs, which is innovative in the field.

Given this and the immense amount of data used in its training – 11 million images and over 1 billion masks [1] – many researchers have recognized the potential of this technology in the medical field and have begun to investigate its effectiveness in this area. However, despite having this large volume of data in its training, there are no medical images among the domains in which SAM was trained, which makes its generalization ability moderate when it comes to this area [2], [3].

This study aims to advance the application of SAM in the field of medical image analysis, especially, for lung segmentation in chest X-ray images. Understanding the effectiveness of SAM in this domain is of paramount importance in the development of new technologies for the diagnosis, treatment and follow-up of lung diseases. We finetune SAM on two collections of chest X-ray images, known as the Montgomery and Shenzhen datasets [4]. Our exploration also involved testing SAM across such datasets using various input prompts, like bounding boxes and individual points. The obtained results show that our finetuned SAM can perform similar to state-of-the-art approaches for lung segmentation, like U-Net [5].

II. Related work

Several studies have conducted a comprehensive evaluation of SAM on a variety of medical image segmentation tasks [2], [3], demonstrating that the model achieved satisfactory segmentation results, especially on targets with well-defined boundaries. However, it is evident that SAM has certain limitations due to the lack of contour in the regions of the images in question, making it difficult to identify certain patterns such as the shapes of organs and tissues [3].

Among such studies, it is worth mentioning the work of Ma and Wang [6]. They introduce *MedSAM*, which was developed on an unprecedented set of over 1 million medical image-mask pairs. Also, they evaluated the fine-tuning of the model, the same technique adopted in this work for chest X-ray images, and the obtained results demonstrate the great potential of SAM in medicine. In most of the 86 tasks evaluated, *MedSAM* ranked first, surpassing the performance of specialized models, like U-Net [5].

Despite the advances, SAM achieved a maximum F1-Score of 60% for lung segmentation in chest X-ray images using points as input, indicating poor performance, as highlighted in the work of He et al. [7]. This study also revealed that the performance with bounding box prompts was even worse.

III. Methodology

A. Technical Approach

The approach used to improve SAM's performance, as previously mentioned, was fine-tuning. After conducting tests with different prompts for the adjustment, including bounding boxes, points extracted from average images, and a combination of both, it was found that the most effective approach was using sets of points obtained from the average images of each dataset. This is due to their better performance.

It is important to highlight that, to avoid any interference in the validation and test sets, only the samples designated for training were used in constructing such images and points. This ensures an objective evaluation of SAM's predictions.

For the bounding boxes needed for training, a function was used to extract the boxes from the image masks and apply perturbations to their coordinates to improve the model's robustness and generalization.

During the fine-tuning process, it was necessary to ensure that gradients were calculated exclusively in the mask decoder, to avoid undesired changes in the other components of SAM. This approach aims to preserve the representations learned in the upper layers of the neural network, which may contain valuable information for the segmentation process.

B. Evaluation Metrics

1) Intersection over Union (IoU): In the *Intersection over Union* metric, the similarity between the mask produced by the model (M_p) and the provided ground truth mask (M_v) is evaluated. This is calculated as the ratio between the intersection and the union of the two masks:

$$\text{IoU}(M_v, M_p) = \frac{|M_v \cap M_p|}{|M_v \cup M_p|},$$

where $M_v \cap M_p$ and $M_v \cup M_p$ are, respectively, the intersection and union between sets M_v and M_p, and $|\cdot|$ is their cardinality.

2) F1-Score: The F1-Score, with a purpose similar to the previous metric, is an evaluation measure that combines precision and recall into a single number, governed by the following expression:

$$\text{F1-Score}(M_v, M_p) = \frac{2 \cdot |M_v \cap M_p|}{|M_v| + |M_p|}.$$

C. Training Procedure

Before starting the experiments, the datasets were divided into validation, training, and test sets. This division is a common practice in machine learning to evaluate and adjust the training and applied techniques. In this work, 20% of the data was set aside for testing, 60% for training, and 20% for validation. The validation set was used to adjust some hyperparameters during the process, such as the segmentation threshold and the loss function.

For comparison with the results obtained with the U-Net network [5], a 5-fold cross-validation was performed in the same manner as the work of Brioso [8]. Cross-validation is a technique to assess the generalization capability of the model on a given dataset. In short, cross-validation splits the data into several subsets called folds. After this procedure, the model is trained, tested, and validated with different portions of the data until all parts have been used as test sets. At the end of the process, the evaluation metrics, such as F1-Score and IoU, are aggregated from all the training and testing iterations to provide a more robust and less biased estimate of performance.

D. Loss Function and Optimizer

Regarding the loss function, the "*DiceFocalLoss*" from the Monai library[1] was chosen, which computes both Dice and Focal losses and returns a weighted sum of them. This approach demonstrated superior performance compared to other available loss functions for image segmentation, such as "*DiceCELoss*", which combines Dice and Cross Entropy losses. As for the optimizer, Adam was chosen, a widely-used method that combines the benefits of RMSProp and SGD Momentum to converge quickly to the global minimum.

IV. Experiments and Results

This section presents the experiments conducted to evaluate and optimize SAM's performance. Initially, the results of the model before fine-tuning are discussed, followed by an analysis of the learning curve. Subsequently, we explore the influence of the segmentation threshold on the quality of the masks generated by SAM and detail the hyperparameter tuning process that optimized the model's performance. Finally, experiments about the model's adaptability and its comparison with other baselines are discussed.

A. Initial Evaluation of SAM

Before fine-tuning SAM for lung segmentation in chest X-ray images, an initial evaluation was conducted to measure the performance of the neural network in its original form [1]. This step is important to establish a baseline reference for the model with respect to the Montgomery and Shenzhen datasets [4]. For this, the datasets were divided into five subsets to follow the 5-fold cross-validation procedure. Table I presents the mean and standard deviation of the evaluation metrics calculated on the resulting subsets obtained from the Montgomery and Shenzhen datasets, respectively.

TABLE I
Initial evaluation of SAM on each of the datasets.

Dataset	Prompt	F1-Score	IoU
Montgomery	**Bounding Box**	0.718 ± 0.033	0.586 ± 0.033
	Points	0.860 ± 0.013	0.774 ± 0.018
	Both	0.848 ± 0.006	0.746 ± 0.007
Shenzhen	**Bounding Box**	0.782 ± 0.009	0.661 ± 0.013
	Points	0.726 ± 0.021	0.593 ± 0.026
	Both	0.863 ± 0.005	0.765 ± 0.008

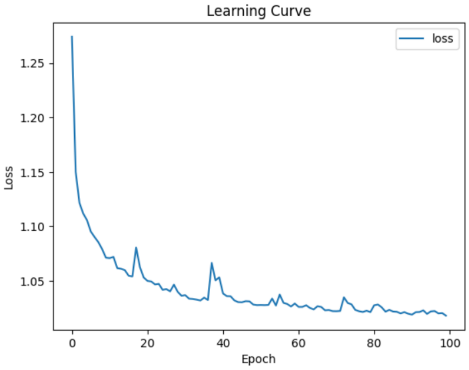
B. Learning Curve

Initially, to determine the appropriate number of epochs for fine-tuning SAM, the learning curve was used. This curve represents the average loss obtained by the neural network over the training epochs and is an essential tool for evaluating its progress and performance. The learning curve allows analysis, through the decrease in loss over time, of whether the model is optimizing over the epochs or stagnating, the latter indicating that the training can be stopped.

From the tests, the conclusion was that none of the datasets and none of the different prompts evaluated showed a significant improvement in loss after the hundredth epoch. Fig. 1 shows an example of one of the learning curves obtained on the Shenzhen dataset.

[1] https://monai.io/ (As of July, 2024)

Fig. 1. Example of a learning curve.



C. Segmentation Threshold

Another critical factor influencing SAM's performance after fine-tuning is the segmentation threshold used to convert the segmented mask into a binary mask during prediction. This is because the mask returned by the model has continuous values between 0 and 1, representing the probability of a given pixel belonging to the region of interest. The choice of threshold to binarize this mask directly affects the results, as an inappropriate selection of this value can lead to inaccurate or incomplete segmentation masks. Low values (e.g., below 0.5) may include many pixels that do not belong to the mask, while high values (e.g., above 0.7) may exclude many pixels that belong to the mask. For this reason, in this study, threshold values from 0.50 to 0.70 were evaluated on the validation set.

The results obtained for the Montgomery and Shenzhen datasets, respectively, are presented in Table II. The best F1-Score for each dataset and prompt is highlighted in bold. This information helped identify the most suitable threshold value for each dataset and input, thereby improving the overall accuracy and quality of predictions.

TABLE II
F1-SCORE OBTAINED FOR DIFFERENT THRESHOLDS IN EACH DATASET.

Dataset	**Threshold**	**Bounding Box**	**Points**	**Both**
Montgomery	**0.50**	**0.828**	0.898	**0.894**
	0.55	0.812	0.932	0.893
	0.60	0.743	0.957	0.879
	0.65	0.596	**0.960**	0.835
	0.70	0.410	0.938	0.759
Shenzhen	**0.50**	**0.837**	0.880	**0.846**
	0.55	0.127	**0.918**	0.782
	0.60	0.003	0.870	0.417
	0.65	0.000	0.829	0.105
	0.70	0.000	0.789	0.005

D. Hyperparameter Tuning

Hyperparameter tuning is a crucial step in the SAM fine-tuning process as it determines the values of learning rate and weight decay, which directly impact the model's performance. Briefly, the learning rate controls the size of steps that the optimization algorithm takes during the learning process. A too high value can lead to instability and prevent or hinder algorithm convergence, whereas a too low value can result in slow or stagnant training. On the other hand, weight decay controls the magnitude of the model's weights by adding a penalty to prevent them from becoming too large, which can help prevent overfitting.

In this study, a grid of predefined values was used: [1×10^{-5}, 1×10^{-4}, 1×10^{-3}] for the learning rate and [0, 1×10^{-1}, 1×10^{-3}] for the weight decay. After the search, a learning rate of 1×10^{-5} and a weight decay of 0 were adopted.

E. Model Adaptability

To verify the model's adaptability, two experiments were conducted in which SAM was trained on one dataset and tested on another. This procedure allows for evaluating the model's generalization capability in different contexts, which is important for ensuring utility and effectiveness in real-world situations where data may vary. It is worth noting that, unlike the other tests, this was performed only with the points obtained from the average image.

First, fine-tuning was performed on the Shenzhen dataset and then applied to the Montgomery dataset. The results of this experiment can be seen in Table III. For a comparison of the results, the F1-Score and IoU values for the Montgomery dataset trained with its own images were also included.

TABLE III
ADAPTABILITY OF THE MODEL TRAINED ON THE SHENZHEN DATASET TO MONTGOMERY.

Metric	**Shenzhen → Montgomery**	**Montgomery → Montgomery**
F1-Score	0.924	0.943
IoU	0.860	0.897

Similarly, another experiment was conducted where the neural network was initially fine-tuned on the Montgomery dataset and then applied to the Shenzhen dataset. Interestingly, the evaluated results were better compared to the fine-tuning done directly on the Shenzhen dataset, as shown in Table IV.

TABLE IV
ADAPTABILITY OF THE MODEL TRAINED ON THE MONTGOMERY DATASET TO SHENZHEN.

Metric	**Montgomery → Shenzhen**	**Shenzhen → Shenzhen**
F1-Score	0.933	0.915
IoU	0.875	0.845

F. Best Models

Table V presents the results of the best models for the Montgomery and Shenzhen datasets, respectively. They were trained using the best hyperparameters found in the previously conducted experiments and with different prompts. The results were obtained based on 5-fold cross-validation, as described in Section III-C. To measure the variability of the results with respect to the mean, the standard deviation is also indicated.

TABLE V
DETAILED RESULTS ON EACH OF THE DATASETS.

Dataset	Prompt	F1-Score	IoU
Montgomery	Bounding Box	0.818 ± 0.040	0.707 ± 0.045
	Points	0.943 ± 0.007	0.897 ± 0.012
	Both	0.876 ± 0.015	0.787 ± 0.023
Shenzhen	Bounding Box	0.797 ± 0.014	0.667 ± 0.024
	Points	0.915 ± 0.011	0.845 ± 0.018
	Both	0.845 ± 0.012	0.735 ± 0.018

G. Comparison with U-Net

A comparison with the best results reported by Brioso [8] for the U-Net network [5] is also performed. The values used for comparison with U-Net represent the best performance achieved by SAM on the test set. These results correspond to the models trained with points as input.

TABLE VI
SAM COMPARED TO U-NET (F1-SCORE).

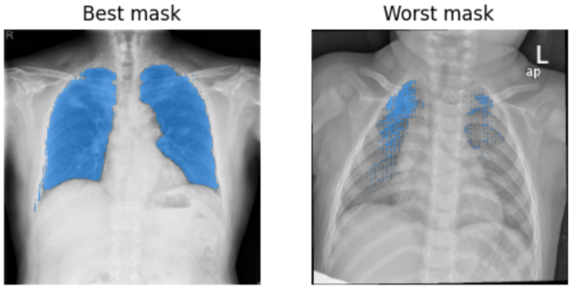
Dataset	SAM	U-Net [8]
Montgomery	0.943 ± 0.007	0.973 ± 0.014
Shenzhen	0.915 ± 0.011	0.941 ± 0.047

Above, in Table VI, F1-Score metric values are compared between SAM and U-Net, as presented in [8]. The results indicate that the U-Net network, known for its effectiveness in this task, achieved superior performance on both the Montgomery and Shenzhen datasets. Unlike the findings of He et al. [7], despite a slightly lower F1-Score, SAM demonstrated satisfactory results similar to those of U-Net, indicating its feasibility and potential for future applications.

H. Predicted Masks

Finally, for illustration purposes, Fig. 2 shows the best and worst predictions made by the model on the Shenzhen dataset.

Fig. 2. Best and worst masks obtained on the Shenzhen dataset, respectively.



V. CONCLUSION

In light of the promising neural network proposed by Meta AI, this work sought ways to fine-tune SAM and adapt it for lung segmentation in chest X-ray images provided by the Shenzhen and Montgomery datasets [4]. Its main objective is to contribute to the automation and accuracy of medical diagnosis, as well as to provide insights and opportunities regarding this new technology.

Throughout this study, various strategies to fine-tuning SAM were investigated, including the use of different prompts, such as bounding boxes, points selected from the average image, and a combination of both. Additionally, different numbers of training epochs and loss functions were explored, with the latter being a crucial choice for satisfactory results.

The experiments conducted during the mask prediction phase allowed for evaluating the impact of the threshold used in the binarization process, that is, converting the masks from the soft mask format to the hard mask format. This analysis provided valuable information on the model's sensitivity to different thresholds and their significant influence on mask quality. Furthermore, tests conducted to verify the model's adaptability showed its utility in real-world situations where datasets exhibit variations.

Considering the complexity of the challenging task of segmenting lungs in chest X-ray images, future work can explore different transfer learning techniques not covered in this study to further improve the model's performance.

ACKNOWLEDGMENTS

This research was supported by São Paulo Research Foundation - FAPESP (grant 2023/17577-0) and Brazilian National Council for Scientific and Technological Development - CNPq (grants 315220/2023-6, 420442/2023-5, and 146570/2023-5).

REFERENCES

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick, "Segment anything," *CoRR*, vol. abs/2304.02643, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.02643

[2] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, S. Liu, H. Chi, X. Hu, K. Yue, L. Li, V. Grau, D. Fan, F. Dong, and D. Ni, "Segment anything model for medical images?" *Medical Image Anal.*, vol. 92, p. 103061, 2024. [Online]. Available: https://doi.org/10.1016/j.media.2023.103061

[3] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Anal.*, vol. 89, p. 102918, 2023. [Online]. Available: https://doi.org/10.1016/j.media.2023.102918

[4] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant Imaging Med Surg*, vol. 4, no. 6, pp. 475–477, Dec 2014.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[6] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024. [Online]. Available: https://doi.org/10.1038/s41467-024-44824-z

[7] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, "Accuracy of segment-anything model (SAM) in medical image segmentation tasks," *CoRR*, vol. abs/2304.09324, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.09324

[8] E. R. C. Q. Brioso, "Anatomical segmentation in automated chest radiography screening," Ph.D. dissertation, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, July 2022, outras ciências da engenharia e tecnologias, openAccess. [Online]. Available: https://hdl.handle.net/10216/143015