Self-supervised 3D Anatomy Segmentation Using Self-distilled Masked Image Transformer (SMIT)

Elycheva DRAY, Théo DI PIAZZA

March 5, 2023

Contents

1. Context, Motivation

Multi-organ 3D segmentation is a very important area of research in medical imaging because when it is done correctly and accurately, it can help for the delineation of certain anatomical structures to assist physicians in surgery, biopsies, clinical tests and it can even aid in tumour detection.

To meet this need, supervised learning methods have already been developed, based on CNNs or vision transformers [2], [3]. These methods consist in training a backbone to predict for each image's pixel, the associated organ. During the course, we also talked about segmentation methods by sliding window classification, which remain very inefficient because they use only a small part of the image for each prediction. These methods present very satisfactory results but they require a large amount of labeled data to obtain such results. A problem is that labeling process is costly and time consuming, and it can be particularly difficult to obtain such datasets in medical imaging.

Thus, unsupervised learning methods have been developed to teach neural network to learn relevant visual representations without labels. For example, [4] uses the contrastive learning method for medical image segmentation which consists in maximizing the similarities of the representations provided by a backbone from 2 different views of the same image. An alternative to CL, whose negative point is the difficulty to select pairs of similar/dissimilar images, : [5] proposes to use a Mask-RCNN to reconstruct the masked part of the input image before fine-tuning the backbone on segmentation tasks.

2. Methodology, Contributions

That's why SMIT [1] is introduced : its aim is to train a backbone able to learn visual representations from an image in a pre-training task, to then fine-tune it on a 3D image segmentation task. This is achieved by combining the advantages of : self-distilled learning; ViT for the attention mechanism; and Masked Image Prediction (MIP) to be able to learn relevant visual representations without labels.

SMIT uses self-distillation to train a Student network to learn visual representations, with the help of a Teacher network. From a given input image, 2 different views are generated with data augmentation method. The first view is corrupted : some masks are randomly masked before going to the Student, whose goal is to predict 1) masked patches, 2) patch tokens of the masked patch and 3) global image token of the embedding. In parallel, the 2 initial augmented views are given to the Teacher network whose goal is to encode 1) the patch tokens of the masked patch from the first initial augmented view and 2) the global image token obtained from the second initial augmented view.

About this first step, the model is pre-trained on 3643 CT patients, i.e. 602708 images, to predict the missing regions of the input images by minimising the reconstruction error. About Masked Patch Token Self-Distillitation (MPD), it is trained such that the student predicts the tokens of the teacher network. Finally, Global Image Token Self-distillation (ITD) is trained by matching the class token distribution extracted from the corrupted view by the student, with the class token distribution from the uncorrupted and different view by the teacher. The loss consists of 3 terms (MIP, MPD, ITD), details of which are in the appendix.

The second step is a fine-tuning of the student backbone on labeled datasets of CT/MRI scans for abdominal organ segmentation. For CT, it was fine-tuned to generate volumetric segmentation of 13 different abdominal organs with 21 random images for training, others for validation. For MRI, it was fine-tuned using 5-fold CV to segment abdominal organs at risk for pancreatic cancer radiation treatment.

3. Validation, Results

About segmentation accuracy on CT/MRI datasets, SMIT achieves superior segmentation performance while using significantly less labeled data. It has outperformed other SSL methods in terms of segmentation accuracy, with an overall DSC score of 0.848 on CT and 0.875 on MR, and was most accurate for organs with highly variable appearance and size. The paper's main results demonstrate the effectiveness of the SMIT architecture and its ability to handle unlabeled data. On average the model shows a higher accuracy compared to the previous best (+0.015 on CT compared to iBOT [6] and +0.027 compared to iBOT on MR; Table 2, [1]), especially on difficult to segment organs such as small bowel, due to the presence of closely packed bowel loops, with an improvement of +0.016 compared to SSIM [7]; and stomach, difficult to segment due to

its highly variable appearance and size, with an improvment of +0.003. To show the impact and efficiency of MIP for SSL which is the main contribution of the paper, it was shown that SMIT outperforms iBot which also uses MPD and ITD for SSL without MIP, regardless of the size of the dataset used for fine-tuning. Moreover, in the fine-tuning task, pre-training allows to obtain 0.8 segmentation accuracy after about 10 epochs, whereas it would take more than 150 epochs to obtain equivalent accuracy without pre-training, which represents a gain in terms of execution time.

4. Dicussion

Regarding paper's strenghts, the model is tested on the CT test set, and on MRI: a dataset totally different from the training dataset. The method shows robustness with good performances on new patients, i.e. with different organs and images from the training dataset. One limitation may be that the model was only tested on a different dataset from the training one. Even if medical data are difficult to obtain, an idea for improvement would be to test the method on other datasets. Besides, the main contribution of the paper is the dense pixel-wise regression pretext task performed within masked patches called masked image prediction with masked patch token distillation to pre-train ViT, which allows to improve the accuracy on segmentation while reducing the number of data needed for fine-tuning.

Moreover, SMIT method clearly states step by step the procedure : architecture, loss functions and implementation details are given which allows us to easily reproduce the experiments. Also, figures and tables allow to visualise the results by comparing them to the state of the art on segmentation on each organ on CT/MRI datasets. A strong point is that SMIT presents better results on the segmentations of organs with high appearance variability and small sizes, than the previous methods. The validation process highlights in detail the impact of MIP for SSL's contribution, thus supporting the effectiveness of SMIT's contribution. Finally, the paper also has a very good reproducibility because the github of the project is publicly available.

However, there are some weaknesses in the methodology that need to be addressed. One issue is the lack of information about computation time. The paper does not provide any information about the computation time required to train the two encoders, which is a crucial factor to consider when comparing the SMIT model with other methods. Since vision transformers are very complex models we can wonder if it is really worth it to use this method given the small improvement of accuracy. To answer this problem objectively, one idea could be to reproduce these experiments with different pre-training methods such as SMIT and iBot in similar contexts, and calculate the pre-training times for each method, to evaluate the trade-off between complexity and performance gain.
Additionally, we are missing information about patients : the paper does not provide explicit information about the patients, such as age or sex.. which could impact the generability of the results.
Besides, we think that the method presented in the paper may lack a bit of novelty : it obviously proposes a new approach but it does not introduce a fundamentally new concept or method but rather integrate existing techniques into a cohesive framework. While this may be useful, it limits the novelty of the approach.
Finally as the final task is organ segmentation, to deepen the work and get more general results, it could be interesting to evaluate the performance of the model on the segmentation task with different metrics to get a more comprehensive understanding of its performance. Even if DSC is the most popular metric, it might be wise to calculate such metrics as IoU which allows to understand the proportion of overlapping pixels; F1-score which can help to diagnose when the model makes mistakes; or Pixel accuracy, to simply understand how well the model captures the overall shape of organs.

Conclusion In this work, the SMIT method is used to train an encoder to learn relevant visual representations, which can then be used for a segmentation task. This work presents the advantages of the methods used: self-supervised learning; ViT; dense pixel-wise regression pretext task performed within masked patches; masked patch reconstruction; in order to present itself as one of the most efficient methods for segmentation on CT and MRI datasets.

References

[1] J. Jiang, N. Tyagi, K. Tringale, C. Crane, H. Veeraraghavan : Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT) (2022)

[2] Y., Zhang, J., Shen, C., Xia, Y.: COTR: efficiently bridging CNN and transformer for 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (2021)

[3] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2022)

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, Ender Konukoglu : Contrastive learning of global and local features for medical image segmentation with limited annotations (2022)

[5] B. Felfeliyan, A. Hareendranathan, G. Kuntze, D. Cornell, N. D. Forkert, J. L. Jaremko, J. L. Ronsky : Self-Supervised-RCNN for Medical Image Segmentation with Limited Data Annotation (2022)

[6] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong : iBOT: Image BERT Pre-Training with Online Tokenizer (2022)

[7] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, H. Hu : SimMIM: A Simple Framework for Masked Image Modeling (2022)

Appendix

Appendix A : Details about the architecture of the model

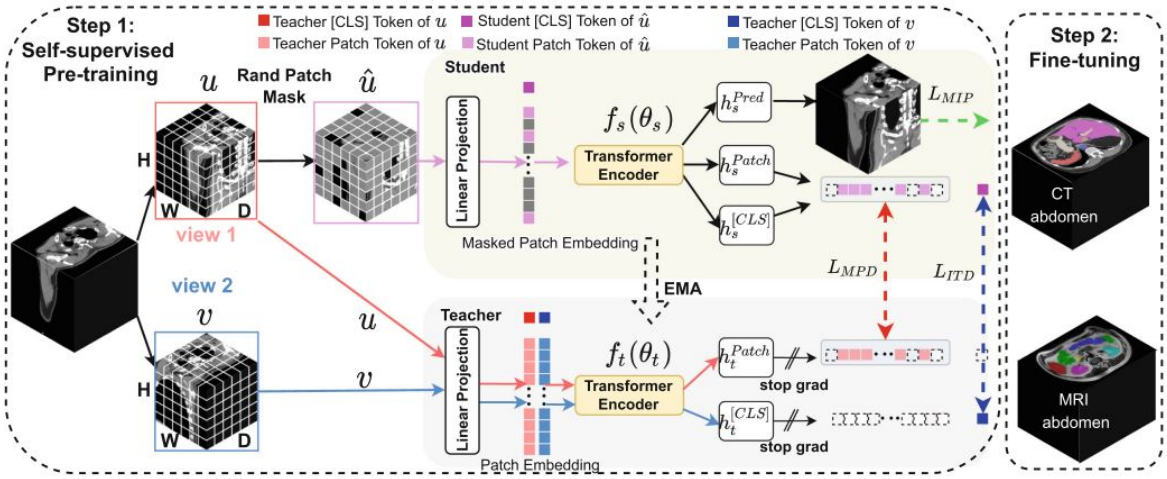

Figure 1: SMIT Scheme. Source : https://arxiv.org/abs/2205.10342

More details about SMIT :

From a 3D image $\mathbf{x}$, 2 views are obtained with data augmentation methods : $\mathbf{u}$ and $\mathbf{v}$. For each view, N images patches are extracted to create a sequence of image tokens. From the first view, some patches are randomly masked to give $\hat{\mathbf{u}}$.

Two networks must be considered : the **TEACHER** and the **STUDENT**.

- The **TEACHER** receives the 2 views $\mathbf{u}$ and $\mathbf{v}$ that are first given into a Linear projection whose output which are **Patch Embedding** are given into the **TEACHER** Transformer Encoder. The encoder outputs $\mathbf{h_t^{Patch}}$ and $\mathbf{h_t^{[CLS]}}$. $\mathbf{h_t^{Patch}}$ corresponds to the patch tokens of the masked patch of the view $\hat{\mathbf{u}}$, obtained from the patch embedding of the view $\mathbf{u}$. $\mathbf{h_t^{[CLS]}}$ corresponds to the global image token obtained from the patch embedding of $\mathbf{v}$.
- The **STUDENT** receives $\hat{\mathbf{u}}$ which corresponds to the view $\mathbf{u}$ with masked patches. It is given into a Linear projection whose output is the corresponding **Masked Patch Embedding** and is given to the **STUDENT** Transformer Encoder. The encoder outputs $\mathbf{h_s^{Patch}}$, $\mathbf{h_s^{[CLS]}}$ and $\mathbf{h_s^{Pred}}$. $\mathbf{h_s^{Patch}}$ corresponds to the patch tokens of the masked patch of the view $\hat{\mathbf{u}}$, obtained from the patch embedding $\hat{\mathbf{u}}$. $\mathbf{h_s^{[CLS]}}$ corresponds to the global image token obtained from the patch embedding of $\hat{\mathbf{u}}$. $\mathbf{h_s^{Pred}}$ corresponds to the prediction of masked patches.

About losses :

- $\mathbf{h_s^{Patch}}$ is compared to $\mathbf{h_t^{Patch}}$: it gives $\mathbf{L_{MPD}}$ i.e. patch tokens of the masked patch from $\hat{\mathbf{u}}$ with the **STUDENT** are compared to the ones obtained from $\mathbf{u}$ with the **TEACHER**.
- $\mathbf{h_s^{Pred}}$ that are the prediction of masked patches from the **STUDENT** is compared to the true patches from the view $\mathbf{u}$: it gives $\mathbf{L_{MIP}}$.
- $\mathbf{h_s^{[CLS]}}$ which is the global image token from $\hat{\mathbf{u}}$ obtained with the **STUDENT** is compared to the global image token from $\mathbf{v}$ with the **TEACHER** network : it gives $\mathbf{L_{ITD}}$.

Appendix B : CT Scan dataset

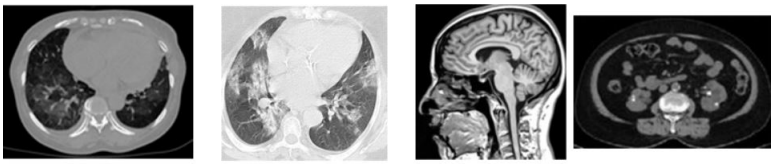

Figure 2: Four 2D images from CT Scan dataset. Source : CT Scan Dataset

Appendix C : Segmentation results

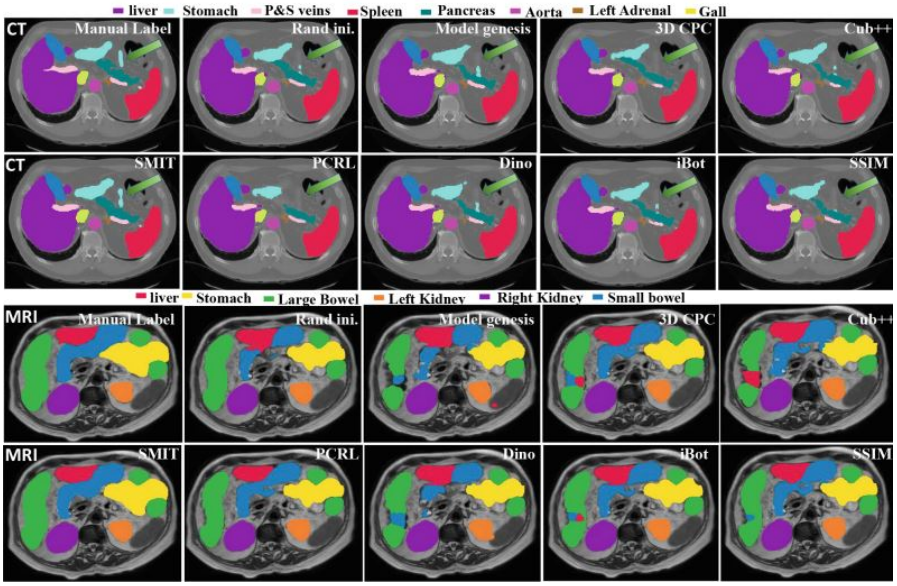

Figure 3: Segmentation results on CIT and MRI datasets. Source : https://arxiv.org/abs/2205.10342