Medical Image Segmentation using text enhanced Vision Transformers

Sindhura Kommu
Virginia Tech
sindhura@vt.edu

Sahana Bhaskar
Virginia Tech
sahanab@vt.edu

Jiayue Lin
Virginia Tech
jiayuelin@vt.edu

Abstract

In the context of clinical images, segmentation plays a pivotal role with a multitude of practical applications. This project aims to enhance the accuracy and efficiency of medical image segmentation tasks by incorporating domain-specific language information. While deep learning has proven invaluable for medical image segmentation, it is often hindered by the scarcity of high-quality labeled data, limiting its performance. To address this challenge, we utilize novel approach that involves training vision transformers for image segmentation while bolstering the model with complementary medical text annotations. The inclusion of text data not only aids in guiding the generation of higher-quality pseudo labels in semi-supervised learning but also leads to an improved fusion of image and text representations. Our evaluation is conducted using 2 multimodal medical segmentation datasets, each consisting of both images and corresponding text data, encompassing X-rays and Tissue Samples. Through a series of experiments, we intend to showcase the findings of our approach and the performance of our model.

1. Introduction

In the realm of clinical images, segmentation assumes a crucial role with myriad practical applications. Despite the transformative impact of deep learning on medical image segmentation, its efficacy is frequently impeded by the scarcity of meticulously labeled data, thereby constraining its overall performance. The precise extraction of the intended object remains a formidable challenge, particularly when dealing with intricately structured target organs characterized by high tissue complexity. Recent investigations underscore the potential of deep learning as a promising avenue for automating medical image segmentation, leveraging the capacity to assimilate and extract the expertise of professionals through specific deep learning methodologies.

Many existing solutions in the field utilize shared encoders, shared decoders, or modality interaction modules [2]. However, the creation of high-quality medical image datasets faces inherent challenges that significantly hinder their application. Obtaining top-tier images is difficult, and the high cost associated with data annotation compounds the problem, imposing constraints on the performance enhancement of medical image segmentation models. Given the complexity of improving both the quantity and quality of medical images, a more practical approach involves leveraging complementary and readily accessible information to compensate for the inherent quality deficiencies in medical images.

Within this context, Picture Archiving and Communication Systems (PACS) emerge as pivotal repositories containing not only medical images but also comprehensive reports generated by radiologists. These reports serve as the official documentation of physicians' interpretations during radiological exams, playing a vital role in conveying findings to patients and healthcare teams. Furthermore, they furnish radiologists with crucial context regarding prior imaging results, particularly during the interpretation of follow-up exams. Radiologists, in reviewing current images, frequently examine prior images and reports to establish the disease's location and extent, facilitating the monitoring of disease evolution and treatment effectiveness. Despite the time-consuming nature of reviewing past exams, its undeniable value in numerous diagnostic applications prompts a shift of focus toward written medical notes accompanied by medical images.

In this work, we used LViT model [2], which is innovative in processing images and text to address the challenge of improving the segmentation performance by using the existing image-text information. In this model, the text feature vector is obtained by using a embedding layer instead of text encoder, which can reduce the number of parameters in the model. In addition, the hybrid CNN-Transformer structure can better merge text information and encode global features with Transformer while retaining the CNN's ability to extract local features from images.

Our contributions include the integration of language cross attention during the reconstruction phase, aligning images with corresponding texts. A language encoder is introduced to map concepts from language space into medical image space, guiding the segmentation process. Additionally, alternative embeddings like BioBERT and ClinicalBERT capture semantic information in clinical notes, enhancing the model's understanding of complex tissue structures.

To further bolster segmentation performance, we incorporate empirically proven vision augmentation techniques and enhance the Dice Loss with Focal Loss during training. The amalgamation of these strategies aims to overcome the challenges posed by limited data quality and quantity, demonstrating the potential for significant advancements in medical image segmentation.

Our experimental validation involves two multimodal medical image segmentation datasets: QaTa-COV19, comprising X-rays, and MoNuSeg, containing tissue samples. The results showcase notable improvements in segmentation accuracy, illustrating the efficacy of our approach. The discussion delves into the implications of our findings, addressing potential biases and highlighting the broader impact on clinical workflows.

In conclusion, our work represents a comprehensive effort to enhance medical image segmentation through the fusion of image and text information. By leveraging existing data sources and innovative deep learning techniques, we aim to contribute to the ongoing evolution of medical imaging for improved patient outcomes.

2. Related Work

Existing language-vision pretraining models, such as Vision Transformers (ViTs), have gained popularity in various computer vision tasks. [1] use ViTs as the image encoder in a U-Net-style architecture for medical segmentation and it captures long-range dependencies in images, making it suitable for the task of image instance segmentation. Other existing language-vision pretraining models mentioned in [2] are CLIP (Contrastive Language-Image Pretraining) and related models, such as ViLT (Vision-Language Transformer), VLT (Vision-Language Transformer), and LAVT (Language-Aware Vision Transformer) are designed for tasks involving text and image information integration. [3] mentions several language-vision pretraining models used in Computer Vision, including CLIP, ALBEF, BLIP, and others.

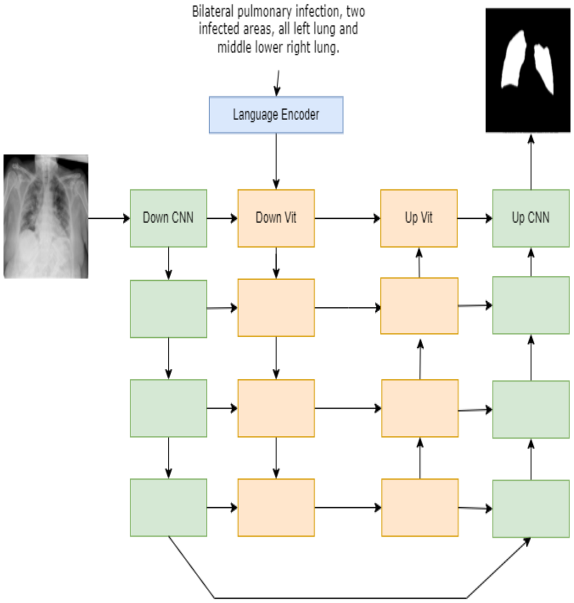
The methodology in [5] involves language-aware visual encoding where language features from a deep language model are combined with visual features through multiple Transformer layers. The model employs a pixel-word attention module (PWAM) to align linguistic meanings with visual cues and a language pathway for controlling the flow of linguistic information.

In [4] the model combines a U-Net architecture for vision feature extraction and a pre-trained language model for text feature extraction. The text embeddings provide semantic information about disease presence and location, guiding the U-Net for precise segmentation. The attention-weighted feature map from this cross-attention process is used for pixel-level prediction.

3. Approach

3.1. Architecture:

The text-enhanced vision transformer model exhibits a dual structure, comprising two distinctive U-shaped branches: a convolutional neural network (CNN) branch and a transformer branch. It also integrates a CNN-ViT interaction module designed to harmonize the features extracted by the Vision Transformer (ViT). Text and Image Vectors are taken as input by the Down VIT branch. The ViT merges the features and harmoniously combines them with CNN features through residual connections, resulting in a CNN-ViT interaction module. The CNN-ViT interaction features are channeled into the UpCNN module to enable the progressive refinement of information layer by layer to generate a mask of the image.


Bilateral pulmonary infection, two infected areas, all left lung and middle lower right lung.
Language Encoder
Down CNN
Down Vit
Up Vit
Up CNN

Language Vision transformer's architecture

3.2. Language Cross Attention

The infusion of language cross attention into the model necessitates structural adjustments aimed at seamlessly merging attention mechanisms that capture the intricate interplay between visual and textual information. This harmonization involves combining attention-weighted visual features with the original visual features within the model, specifically executed during the reconstruction stage.

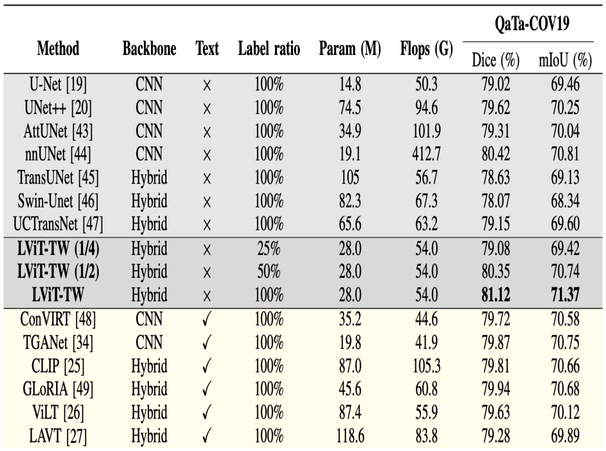
In essence, the text embeddings encapsulate valuable semantic information related to the presence and location of diseases, acting as a crucial guide for the segmentation process. The cross-attention mechanism, operative between the text embeddings and the decoded feature maps, gives rise to a pixel-wise attention map. Following this, the attention map undergoes a normalization process through a Tanh activation function, effectively constraining values within the range of -1 to 1. Subsequently, the normalized pixel-wise attention map is employed in a pixel-wise multiplication with the query feature map. The primary objective of this integration is to augment the model's ability to discern and prioritize pertinent features, fostering a more nuanced and contextually informed approach to image segmentation.

However, the model's performance post-incorporation has not met expectations. This could be attributed either to the implementation approach or the inherent challenge that cross-attention tends to emphasize global features, which may not align with the nuanced requirements of medical images. Medical images often demand a more pronounced emphasis on local features, and the current global focus might be a factor contributing to the suboptimal performance observed. Further investigation and potential adjustments are warranted to address these challenges and refine the model for enhanced medical image segmentation.

4. Experiments

4.1. Text Embeddings

Our initial experiments involved using various text embeddings specifically relevant to our problem statement. We focused on Clinical Bert and BioBert, which are specialized adaptations of the BERT (Bidirectional Encoder Representations from Transformers) model, designed for distinct applications in the biomedical and clinical domains. These models yielded superior results, as evidenced by the data presented in the table [2]. The tailored nature of Clinical Bert and BioBert, with their training on domain-specific texts, enabled them to perform more effectively in our context than the general BERT model.



Method	Backbone	Text	Label ratio	Param (M)	Flops (G)	QaTa-COV19	
						Dice (%)	mIoU (%)
U-Net [19]	CNN	×	100%	14.8	50.3	79.02	69.46
UNet++ [20]	CNN	×	100%	74.5	94.6	79.62	70.25
AttUNet [43]	CNN	×	100%	34.9	101.9	79.31	70.04
nnUNet [44]	CNN	×	100%	19.1	412.7	80.42	70.81
TransUNet [45]	Hybrid	×	100%	105	56.7	78.63	69.13
Swin-Unet [46]	Hybrid	×	100%	82.3	67.3	78.07	68.34
UCTransNet [47]	Hybrid	×	100%	65.6	63.2	79.15	69.60
LViT-TW (1/4)	Hybrid	×	25%	28.0	54.0	79.08	69.42
LViT-TW (1/2)	Hybrid	×	50%	28.0	54.0	80.35	70.74
LViT-TW	Hybrid	×	100%	28.0	54.0	**81.12**	**71.37**
ConVIRT [48]	CNN	✓	100%	35.2	44.6	79.72	70.58
TGANet [34]	CNN	✓	100%	19.8	41.9	79.87	70.75
CLIP [25]	Hybrid	✓	100%	87.0	105.3	79.81	70.66
GLoRIA [49]	Hybrid	✓	100%	45.6	60.8	79.94	70.68
ViLT [26]	Hybrid	✓	100%	87.4	55.9	79.63	70.12
LAVT [27]	Hybrid	✓	100%	118.6	83.8	79.28	69.89

4.2. Weighted Dice Loss with Focal Loss Elements

This custom loss function, `WeightedDiceLoss`, integrates aspects from both the Dice Loss and the Focal Loss to handle class imbalance during segmentation training.

Key Variables:

1. `alpha`: Controls the contribution of the Focal Loss, assisting in managing class imbalance.
2. `gamma`: Shapes the loss curve, directing the model's focus toward challenging examples.
3. `weights`: Assigns significance to different classes, balancing their impact on loss computation.

Functionality:

1. **Usage**: Takes model predictions (`logit`) and ground truth labels (`truth`) as inputs.
2. **Operation**:
   (a) Reshapes tensors for computations.
   (b) Applies class weights to address the class imbalance.
   (c) Computes intersection, union, and the Dice coefficient.
   (d) Calculates Focal Loss based on the Dice coefficient.
   (e) Derives the mean Focal Loss across the batch as the final loss value.

We improved the Weighted Dice Loss by integrating Focal Loss, a variant of standard cross-entropy loss. This adjusts weights for accurately and inaccurately classified samples and effectively addresses class imbalance.

The formula for focal loss is shown below:

$$\text{Focal Loss} = \alpha \times (1 - P_t)^{\gamma} \times \text{Dice Coefficient}$$

The probability that the model predicts for the ground truth object is denoted as P_t. The parameters α and γ represent the weights and the curve's shape, respectively. γ governs the loss curve's shape; higher values decrease the loss for well-classified examples, extending the range of low loss. When $\gamma = 0$, the equation mirrors Cross Entropy Loss. α addresses class imbalance by assigning higher weights to rare classes and lower weights to dominant or common classes.

This loss function provides a balanced approach to tackle class imbalance during segmentation training. Fine-tuning `alpha`, `gamma`, and `weights` allows customization based on dataset specifics and class importance, thereby aiding in enhancing model performance.

$$\mathbf{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

5. Results

We utilized the Dice coefficient to quantitatively evaluate our vision transformer model's accuracy and precision in performing semantic segmentation on medical images. In addition to this we used the Intersection over Union (IOU) to measure the overlap between predicted and ground truth regions. We used Focal loss to ameliorate the problem of class imbalance in the data. Furthermore, we conducted an ablation study to evaluate the performance of our vision transformer model. This assessment involved the use of 2 multimodal medical image segmentation datasets, incorporating CT images and X-rays from well-established sources such as QaTa-COV19 and MonuSeg, which are widely recognized datasets in the field of medical imaging.

In our experiments with the MoNuSeg dataset utilizing BERT-based-uncased text embeddings, employing Focal Loss instead of Binary Cross-Entropy (BCE) resulted in modest improvements. The use of Focal Loss demonstrated a slight enhancement, achieving 79.18% Dice and 65.92% IoU.

Our findings exhibit enhanced accuracy over baseline models like U-Net and LAVT. However, we have yet to attain parity with the current state-of-the-art model in our domain.

Dataset	Text Embedding	Loss	Dice (%)	iOU(%)
MonuSeg	Bert	BCE	79.03	65.53
MonuSeg	Bert	Focal Loss	79.18	65.92

In our experimentation with the QaTa-Cov19 dataset, we observed diverse performance metrics corresponding to different text embeddings. Notably, the utilization of ClinicalBERT yielded the most promising outcomes, achieving 82.25% Dice and 73.19% IoU. Our model's performance exhibited improvement upon employing these text embeddings, surpassing established architectures designed for language and vision data—specifically, U-NET, U-NET++, ConVIRT, and LAVT.

Dataset	Text Embedding	Dice (%)	iOU (%)
QaTa-COV19	Bert-base-uncased	81.70	72.89
QaTa-COV19	Biobert	81.98	73.04
QaTa-COV19	Clinicalbert	82.25	73.19

Here is a link to the GitHub repository: Github Repo Link

References

[1] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *arXiv preprint arXiv:2306.15350*, 2023.

[2] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: Language meets vision transformer in medical image segmentation, 2023.

[3] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01, 2023.

[4] Junjie Hu Tyler J. Bradshaw. Zachary Huemann, Xin Tie. Contextual net: A multimodal vision-language model for segmentation of pneumothorax. *arXiv preprint arXiv:2303.01615*, 2023.

[5] Yansong Tang Kai Chen Hengshuang Zhao Philip H.S. Torr Zhao Yang, Jiaqi Wang. Lavt: Language-aware vision transformer for referring image segmentation. *arXiv preprint arXiv:2306.15350*, 2022.