Segmentation of Airway Tree structures using transformer based Swin Unetr

Anurag Singh Rathore[1][0000-0002-5913-4244], Manoj Kumar Ramteke[1][0000-0002-3837-8952]

Mehul Kamboj[1][0000-0003-4500-557X], Sanjeet Sanjay Patil[1][0000-0001-6259-6549]

Keerthiveena B[1][0000-0002-0716-8469]

[1] Indian Institute of Technology Delhi(IITD), Delhi, India
director@admin.iitd.ac.in

Abstract
The advent of deep learning has turned automated biomedical image segmentation for clinical usage into a realizable technique. The delineation of airways in thoracic CT scans is critical in the diagnosis of pulmonary diseases. The CNN-based architectures display poor performance in segmenting peripheral bronchi due to intra-class imbalance among the trachea and bronchi. This paper presents a framework for segmentation of the airway tree structure from thoracic CT scans using a vision transformer-based network. The developed method can extract airways with high topological completeness in a fully-automated fashion for end-to-end deployment in clinical use. The segmentations have been carried out by a vision transformer based Swin Unetr (Swin Unet Transformer) architecture. This architecture takes 3D patches of preprocessed Lung CT scans as input and produces binary segmentations of refined airway tree structures. A dice score of 96.5% was obtained after training the architecture on 250 CT scans for 100 epochs and validating it on 48 CT scans.

Keywords: Medical image segmentation, Vision Transformer, Swin Unet architecture, Complete coordinate patching.

1 Introduction

The delineation of airway tree structure from the thoracic CT scans is a tedious and time-consuming process when done manually. The reason being its complex 3D structure, with tubular branches of varying sizes and orientations [1]. The accurate measurements of these fine-grained pulmonary airway structures are crucial for diagnosis of abnormality in patient with chronic obstructive pulmonary disease. The segmentation of patient specific airways, comprising of trachea, main bronchi, lobar bronchi, distal bronchi and the peripheral bronchi is also necessary in bronchoscopic-assisted surgery.

The CNN-based networks, specifically the U-Net architecture has been applied extensively for the segmentations of airway tree structure. These architectures perform well at segmenting the trachea, main bronchi, lobar bronchi and distal bronchi but fail at efficiently identifying the peripheral bronchi. This produces a lot of false positives and false negatives in the automatically segmented labels. Recently, vision transformers have become more ubiquitous due to their ability to access global and local representation simultaneously. Its self-attention module achieves this by processing information using token embeddings for pairwise interactions [2]. The greater receptive field in vision transformers gives it an edge over the CNNs in the task of identifying and segmenting larger and dispersed targets. In this case, we have the airway tree structure which covers almost entire area of lungs. The Swin Unet Transformers (brats reference) are one of those vision transformers employing shifted windows partitioning for computing self-attention in a U-shaped network.

Our framework comprises of 3 main phases: 1) Pre-processing the thoracic CT scans to produce bounded patches of lungs, 2) Segmentation of those patches using Swin-Unetr architecture to produce binary segmentations of airway tree structure, and 3) Post processing of the patches to produce final labels of airways.

2 Materials and Methods

2.1 Preprocessing

The full-size thoracic CT scans are cropped to produce bounding boxes around the lung fields with a buffer of extra 30 voxels from the lung boundaries. This was done to facilitate the architecture's focus on region of interest while the buffer voxels alleviated the effects of boundaries in prediction of peripheral airways. 3D patches

(128*128*128) were generated out of the cropped scans using a forward-reverse complete coordinate patching algorithm to cover entire bounded box.

2.2 Architecture

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.

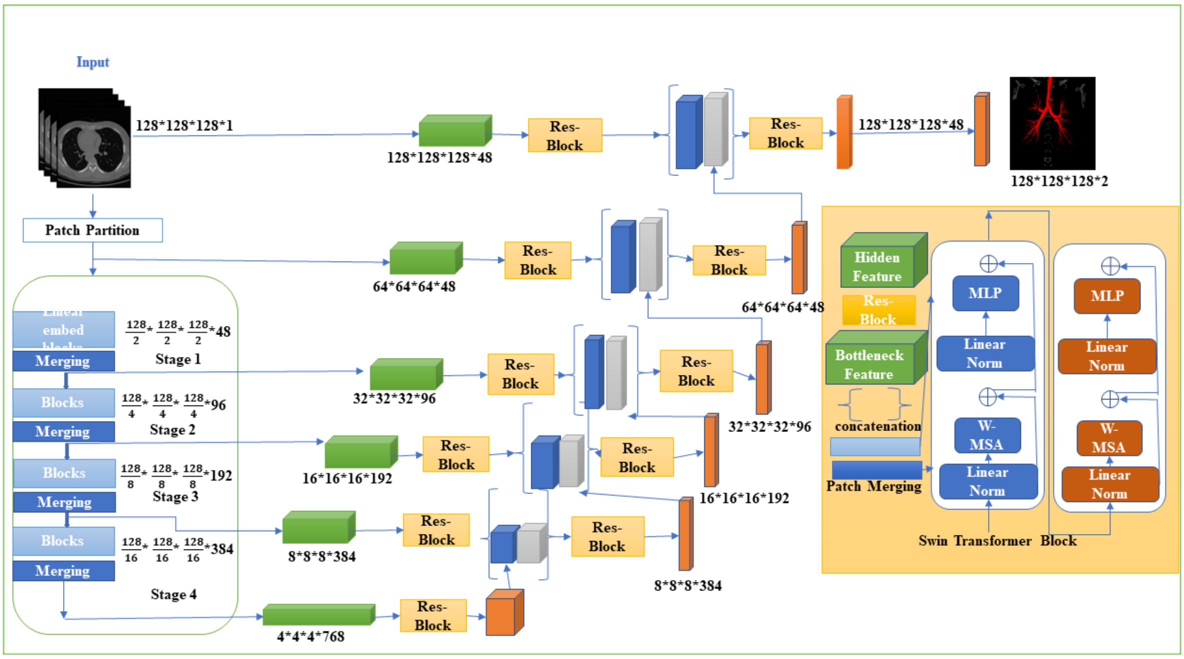

Figure 1 :- Schematic Diagram of Swin Unetr Architecture

The Swin Unetr architecture comprises of a transformer based encoder and a CNN based decoder. The schematic diagram of our Swin-Unetr architecture is given in fig. 1. This architecture was implemented by Hatamizadeh et. al., (2022) to segment 14 abdominal organs from CT scans [3, 4]. The encoder takes pre-processed 3D volumes as input and employs a patch partitioning layer to produce a sequence of 3D tokens which are projected into a 48-dimensional embedding space. These embeddings enter the Swin Transformer Block. The Swin Transformer Block consists of Layer Norm (LN), multi-head self-attention module, residual connection and 2- layer MLP with GELU non-linearity. The regular window based MSA and the shifted window based MSA are applied successively in these blocks.

The encoder of Swin Unetr has a feature dimension off $2\times 2\times 2\ \times 1 = 8$, given the patch size of $2\times 2\times 2$ and single channel 3D input volume. The four stages of encoders have $\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}, \frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}, \frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}$ and $\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}$ resolutions respectively. The patch merging layer is utilized to down-sample the resolution after every stage in the encoder. Moreover, the layer also groups $2\times 2\times 2$ patches and concatenates them, resulting in a 4C dimensional feature embedding. The 4C dimensional embeddings are halved using a linear layer.

The Decoder of Swin Unetr resembles the Unet architecture's shape and structure. It receives the extracted feature representations from the encoder through the skip connections at each resolution. The output feature representations are reshaped and fed into residual blocks comprising of two 3×3×3 CNN. These layers are normalized using instance normalization. The features maps from residual blocks are upsampled by a factor 2 using deconvolutional layers and the outputs from the bottom stage are concatenated with the upsampled outputs. The features from these operations are again fed into residual blocks. In the final layer a 1×1×1 convolutional layer and a sigmoid activation produce the segmentation ouput.

2.3 Loss function, Optimizers and Evaluation metric

The objective of training is to reduce the binary categorical cross entropy loss function. The optimization of the training is done using adam optimizer algorithm. We have used dice similarity coefficient as a metric to evaluate Swin Unetr's training efficiency.

2.4 Implementation Details

The training of our Swin Unetr architecture is done on Nvidia RTX a5000 GPU. The code is implemented using PyTorch and has a backend of MONAI. Random patches of 128×128×128 were cropped out of the pre-processed 3D volumes during training. The cropping of 128× 128 × 128 patches were done randomly. While the decision to keep the centre of the patches as foreground or background voxel was taken according to the positive to negative ratio (pos/neg =1). The architecture has been trained on 300 thoracic CT scans provided in the Airway Tree Modelling challenge [5, 6, 7, 8]. The dataset was normalized between 0 and 1 before feeding it to the architecture and no other augmentations were applied on the dataset. The learning rate during training is set at 10^{-4} and the model was trained for 100 epochs.

3 Results

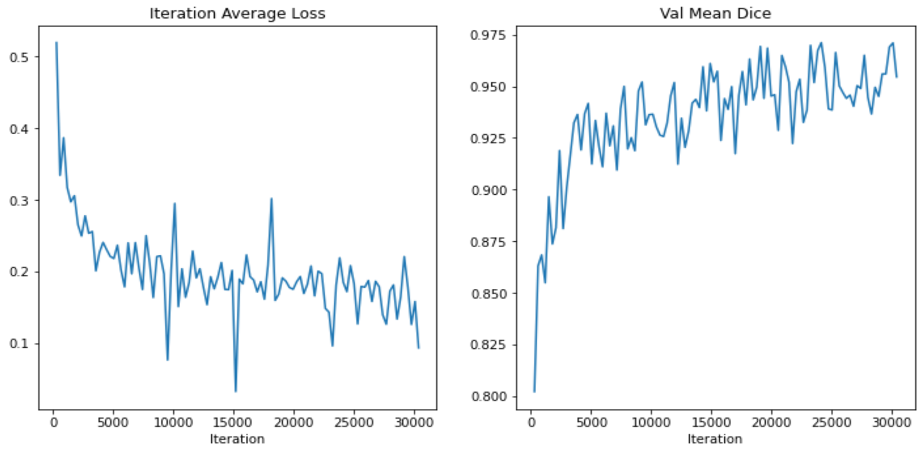

Figure 2:- Iteration Average Loss vs Iteration and Val Mean Dice vs Iteration during training

The training set was split into 250 training images and 48 validation images. After training for over 100 epochs we have achieved a dice score 0.9715 and a loss of 0.1020. The change in training dice score and loss function with the iterations are given in fig. 2. The test dataset was converted into patches of 128×128×128 and given as input to the model to produce segmented patches. These patches were unpatched to produce labels of input image size. The results were then post processed using largest connected components to produce a largest 26 neighbor connected component as the final airway segmentation. The 3D representation of the segmented labels, before post processing and after post processing, on one of the validation dataset provided by the challenge organizers is given in fig. 3. The 3 D visualization was done using the package provided by S. Bakas et al. (2017) [9, 10].

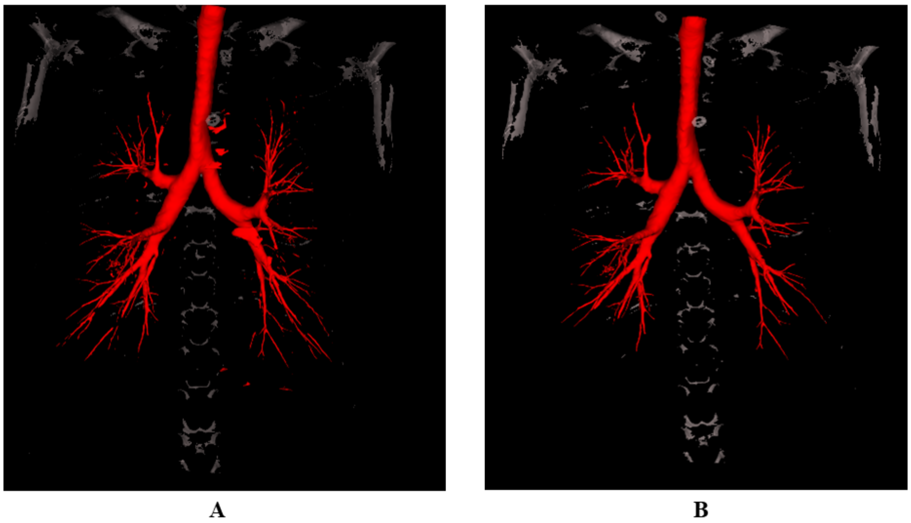

Figure 3:- Comparison of the segmentation results between A:before postprocessing and B:after applying largest connected components postprocessing.

4 Conclusion

This paper comprises of an end to end framework for the segmentation of lung airways from the thoracic CT scans using a transformer based architecture. The patches of the training dataset for training and testing are — created with complete coordinate patching algorithm. This patching mechanism enhanced the architectures attention on fine bronchial branches. We have achieved a dice score of 0.9715 during our training.

References

1] Garcia-Uceda, Antonio, et al. "Automatic airway segmentation from computed tomography using robust and efficient 3-D convolutional neural networks." *Scientific Reports* 11.1 (2021): 1-15.

2] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

3] Hatamizadeh, Ali, et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images." *International MICCAI Brainlesion Workshop*. Springer, Cham, 2022.

4] Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

5] Zhang, Minghui, et al. "Fda: Feature decomposition and aggregation for robust airway segmentation." *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, Cham, 2021. 25-34.

6] Zheng, Hao, et al. "Alleviating class-wise gradient imbalance for pulmonary airway segmentation." *IEEE Transactions on Medical Imaging* 40.9 (2021): 2452-2462.

7] Yu, Weihao, et al. "BREAK: Bronchi Reconstruction by gEodesic transformation And sKeleton embedding." *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022.

8] Qin, Yulei, et al. "Airwaynet: a voxel-connectivity aware approach for accurate airway segmentation using convolutional neural networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2019.

9] Bakas, Spyridon, et al. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The cancer imaging archive." *Nat Sci Data* 4 (2017): 170117.

10] Menze, Jakab, Himmelreich Masuch, and Petrich Bachert. "Menze BH." *Kelm BM, Masuch R., Himmelreich U., Bachert P., Petrich W., et al., A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics* 10.1 (2009).