Deep learning approaches for lesion segmentation on 18 FDG PET/CT imaging[⋆]

Ziwei Zhu[1,2]

[1] Gordon Center for Medical Imaging, Massachusetts General Hospital (MGH) and Harvard Medical School (HMS), Boston, MA, USA
[2] South China University of Technology , Guangzhou, China

Abstract. We used Restormer network architecture for the segmentation of lesions of PET/CT.

Keywords: Segmentation · nnU-Net · Transformer

1 Introduction

1.1 Background

Positron emission tomography/computed tomography (PET/CT) is an integral part of the diagnostic effort for various malignant solid tumor entities. Because of its broad applicability, fluorodeoxyglucose (FDG) is the most widely used PET tracer in the tumor setting to reflect tissue glucose consumption, for example, typically increasing glucose consumption in tumor lesions. As part of the routine clinical analysis, PET/CT is mainly analyzed qualitatively by experienced medical imaging experts. Additional quantitative evaluation of PET information would allow for more precise and individualized diagnostic decisions. A crucial initial processing step for quantitative PET/CT analysis is a segmentation of tumor lesions enabling accurate feature extraction, tumor characterization, oncologic staging and image-based therapy response assessment. Manual lesion segmentation is, however, associated with enormous effort and cost and is thus infeasible in clinical routine. Automation of this task is therefore necessary for widespread clinical implementation of comprehensive PET image analysis. Recent progress in automated PET/CT lesion segmentation using deep learning methods has demonstrated the principle feasibility of this task. However, despite these recent advances, tumor lesion detection and segmentation in whole-body PET/CT is still challenging. The specific difficulty of lesion segmentation in FDG-PET lies in the fact that tumor lesions and healthy organs (e.g. the brain) can have significant FDG uptake; avoiding false positive segmentations can thus be difficult. One bottleneck for automated PET lesion segmentation progress is the limited availability of training data that would allow for algorithm development and optimization.

[⋆] Supported by European Society for hybrid, molecular and translational imaging (ESHI)

In this paper, we aim to explore the potential of using a denoising method based on the SwinTransformer[2]. method to improve the PET images segmentation metric compared with the SOAT medical images segmentation network like nnU-Net[1]. Over the past three years, nnU-Net has been widely and successfully applied to medical images segmentation. Recently, many authors have also used Transformer to improve the nnU-Net to create many variant nnU-Net like nnFormer.

Different from existing images segmentation network based on CNN methods, we use a

2 Method

2.1 Dataset

The training cohort consists of 900 patients(1014 studies) with histologically proven malignant melanoma, lymphoma or lung cancer and negative control patients whom FDG-PET/CT examined in two large medical centers (University Hospital Tübingen, Germany University Hospital of the LMU in Munich, Germany). All PET/CT data within this challenge have been acquired on state-of-the-art PET/CT scanners (Siemens Biograph mCT, mCT Flow and Biograph 64, GE Discovery 690) using standardized protocols following international guidelines. CT and PET data are provided as 3D volumes consisting of stacks of axial slices. Data provided as part of this challenge consists of whole-body examinations. Usually, the scan range of these examinations extends from the skull base to the mid-thigh level. If clinically relevant, scans can be extended to cover the entire body, including the entire head and legs/feet.

2.2 Network Architecture

We add the block Gated Convolution Feedforward Network (GCFN) in SwinTransformer with Restormer[3]. The achitecture shown as Fig.1.
1. The input image size is $I \in R^{H \times W \times 3}$ first, use a convolution operation to obtain the feature embedding $F_0 \in R^{H \times W \times C}$. F_0 obtains a high-dimensional feature F through a symmetrical 4-layer encoding-decoding structure. Each layer of encoding/decoding includes multiple Transformer modules. From top to bottom, the number of Transformer modules in each layer is increased. Resolution gradually decreases. Skip links are used between encoder-decoder to transfer low-dimensional feature information. F_0 further go through the Refinement module to extract detailed features. Finally, go through a volume base layer and superimpose it with the input image to obtain the final output image.

2.3 Data preprocessing

Images Croping Images cropping is to crop a three-dimensional medical image to its non-zero area. The specific method is to find a minimum three-dimensional

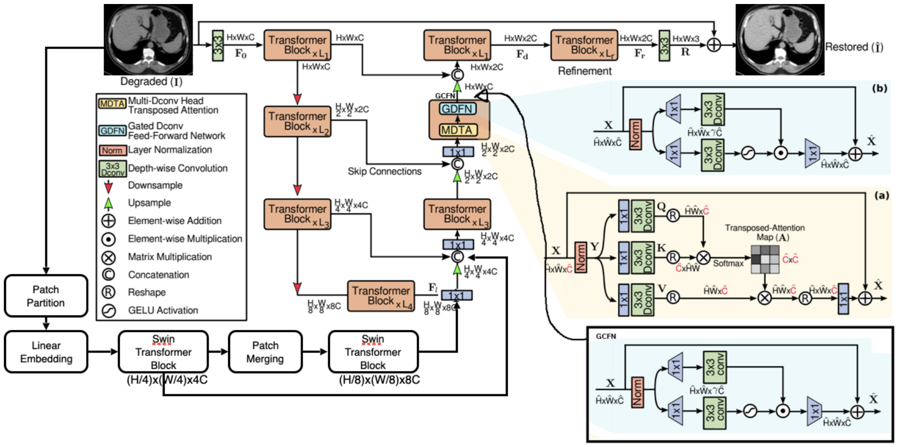

Fig. 1. Swin Transformer + Restormer + CGFN

bounding box in the image. The value outside the bounding box area is 0, and the image is cropped using this bounding box. Compared with before cropping, the cropped image has no effect on the final segmentation result, but reduces the size of the image, avoids useless calculations, and improves computational efficiency. This operation is more effective for CT datasets, with relatively many backgrounds with an all-black periphery. Cutting is divided into three steps. The first step is generating a three-dimensional non-zero template non-zero mask according to the four-dimensional image data (C, X, Y, Z), indicating which areas in the image are non-zero. Different modalities have corresponding 3D data, resulting in different 3D non-zero masks. The non-zero template of the entire 4D image is the union of the non-zero templates of each modality.The second step determines the size and position of the bounding box for cropping based on the generated non-zero template. The third step is to crop each image modality according to the bounding box and then reassemble them together. After cropping the original data, use the same bounding box to crop the mask of the segmentation annotation, and the specific steps are the same as the above steps.

Images Resampling The purpose of resampling is to solve the problem that the actual spatial size represented by a single voxel in different images is inconsistent in some 3D medical image datasets. Because the convolutional neural network only operates in the voxel space, it ignores the size information in the actual physical space. In order to avoid this difference, it is necessary to resize different image data in the voxel space to ensure that in different image data, the actual physical space represented by each voxel is consistent. The first step is to determine the size of the target space for resampling. The second step determines the target size of each image based on target spacing. For each image, the product between spacing and shape is a fixed value, representing the size of the entire image in actual space. The third step is to resize each image.

Images Normalization The preprocessing of normalization is performed on the image so that the grey value of each image in the training set can have the same distribution. The CT image normalization uses the mean and standard deviation of the foreground of the entire training set. It clips the HU value of the image to the percentage range of the foreground HU value [0.5, 99.5]. In PET image normalization, only the grayscale information of a single image is used to calculate the mean and variance.

2.4 Training

Loss Function We train our networks with a comination of dice and cross-entropy loss:

$$\mathcal{L}_{total} = \mathcal{L}_{dice} + \mathcal{L}_{CE}$$

The dice loss is implemented as follows:

$$\mathcal{L}_{\text{dc}} = -\tfrac{2}{|K|}\sum_{k\in K}\frac{\sum_{i\in I}u_i^k v_i^k}{\sum_{i\in I}u_i^k+\sum_{i\in I}v_i^k}$$

Optimizer We use Adam optimizer with learning rate 3e-4; 250 batch/epoch; Learning rate adjustment strategy: Calculate the exponential moving average loss of the training set and the validation set. If the exponential moving average loss of the training set is reduced by less than 5e-3 within 30 epochs, the learning rate will be attenuated by five times; Training stop condition: When the exponential moving average loss of the validation set is not reduced by 5e-3 within 60 epochs, or the learning rate is less than 1e-6, the training would be stopped.

Table 1.

Result	
Metrics	**score**
Dice Score	**0.9247**
False Negative	**1.0052**
False Positive	**2.3563**

3 Result

The Training metrics are list on Table.1. The training curve is shown as Fig.2.The visualization results of the predictions of our method are shown in Fig.3

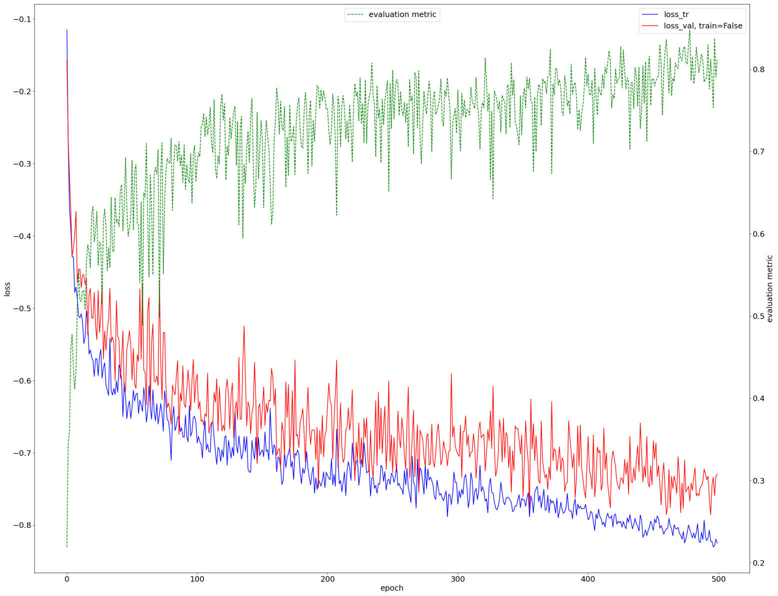

Fig. 2. The Training Curve

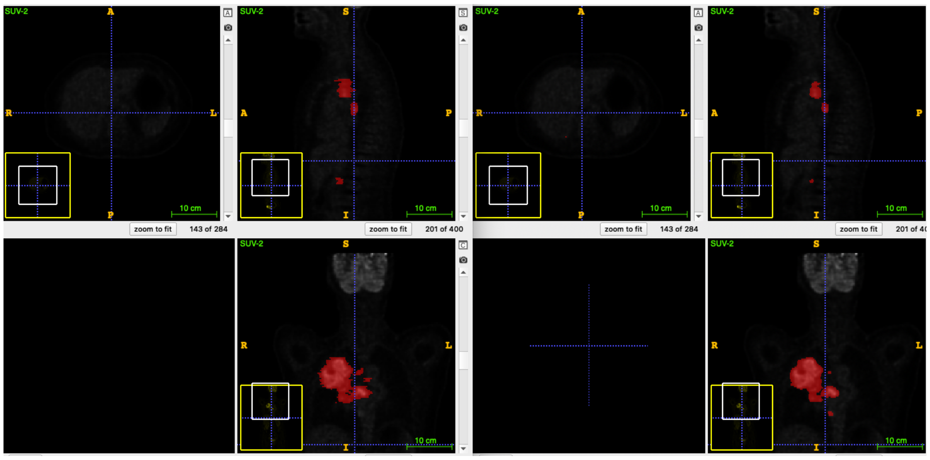

Fig. 3. visualized result left:our right:ground-truth

References

1. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
2. Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–88. Springer, 2021.
3. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.