Applying TransUNet to Renal Tumor Segmentation

Matthew Nguyen, Neel Karsanbhai

Paper ID 2

1 Introduction

1.1 Motivation

The analysis of medical images (MRIs, CT scans, PET images) for cancer diagnosis, and cancer analysis on patients is currently a high level analysis done by human radiologists [1]. This is inefficient from a cost perspective, and data extraction perspective. Radiomics is a practice based on the idea that medical imaging in general can provide more insights into a patient than just visual information. Medical imaging can also provide doctors and other healthcare providers with context. Radiomics can provide correlations between genetics and tissue type, and confirmation of the findings of a radiologist among may other uses [1]. This is a new and promising field of medical imaging that can help tumor boards create more effective and cost efficient treatment plans for their patients. Analysis of this type if leading the way due to support from the National Cancer Institute, and the Quantitative Imaging Network. After collecting image data, and identifying data that is useful, tumors must be identified, and segmented. Automating image segmentation will aid radiologists and largely tumor boards in the process of radiomics, while lowering the costs of cancer care for patients. The current annual cost of cancer treatment in the United States is $109,727 [2]

1.2 Why Current Solutions are Inadequate

The TransUNet architecture proposed by J. Chen et al. showed the efficacy of the architecture on multi-organ abdominal CT scans as well as MR images of the heart. TransUNet achieved superior performance over state of the art architectures (V-Net, DARR, U-Net, and AttnUNet) based on the Dice similarity coefficient [3]. This architecture has been proven to work on multi-organ segmentation, which is within the larger domain of medical image segmentation, however, it has not been proven on the renal cancer segmentation task. Global context and spatial information should be helpful in the renal cancer segmentation task, because both abilities should be able to work together to determine the location of a tumor, as well as the change in tissue type from renal tissue to tumor tissue.

Berbera et al. proposed a network that used three sequential modules and spatial

transformers [4]. The network was able to reduce training time while increasing the Dice score for renal tumor segmentation from 85.52% to 87.12%; however, the model was only valid for a small set of data [5].

1.3 Proposed Idea

Our proposed network will perform renal tumor segmentation on 2D CT scan images using U-Net with visual transformers (ViT), an adaptation of the transformer, proposed by Dosovitskiy et al. [6], that applies global attention to 16x16 patches of an image. The benefit of using this type of transformer is that it is the the best at incorporating global context in the image features without compromising the computational efficiency [5]. Our proposed network will trained and tested using the KiTS19 dataset [7].

2 Related Work

2.1 Combining UNet with ViT for Renal Cancer Segmentation

To the best of out knowledge, this is the first application of ViT and UNet to renal cancer segmentation. This architecture, however, has been shown to outperform the state of the art on multi-organ segmentation. TransUNet itself has performed well on multiple medical applications such as cardiac segmentation, organ segmentation, and polyp segmentation on multiple forms of imaging [3]. The purpose of TransUNet is to overcome the inability of Convolutional Neural Networks (CNN) to model long range relationships. CNNs more specifically have difficulties identifying textures, shapes, and sizes of objects that are highly variable between patients [3]. TransUNet is well suited to medical applications, because it requires less training due to the fact that it uses UNet. UNet is able to do this by up sampling images [8]. By applying the TransUNet architecture to the renal tumor segmentation task, we hope to identify how hyper parameters of the network should change to best fit the task, if changes need to be made at all. This experiment can possibly identify characteristics of the data set that human radiologists may have a difficult time identifying due to bias, inconspicuousness, etc.

2.2 UNet with Spatial Transformer Network for Renal Tumor Segmentation

While the network proposed by Berbera et al. was trained and tested on a dataset of renal CT scans, their dataset is composed of CT scans from pediatric patients. Additionally Berbera et al. used a Spatial Transformer Network (SPN) proposed by Jaderberg et al. [9] while we plan on using visual transformers. SPNs transform image inputs to decrease the noise in an image, locate the area of interest in an image, and increase the invariance of a network [9]. ViT, however applies attention mechanisms to patches of the input image. The data set we will

be using contains abdominal CT scans from adult patients. Berbera et al. uses a normalization layer in addition to a localization layer, and segmentation layer to account for the high variability in kidney size of pediatric patients. Adult patients most likely do not have high inter patient variability with regard to kidney size reducing the need for a normalization layer. Thus our architecture will not include a normalization layer.

3 Methods

3.1 Architecture

Our architecture will exactly replicate the TransUNet architecture proposed by Chen et al. [3].

Input. Our model starts off with an input layer that takes in images with a resolution of 224×224. While results from Chen et al. showed that increasing this input resolution to 512×512 would increase performance, it comes at the expense of a much larger computational cost [3].

CNN. The input images are passed into a CNN that is used as a feature extractor in order to create a feature map. A CNN is used because without it, low-level details would be lost due to the shape of encoded features being much smaller than the original image. Including a CNN, allows us to use the intermediate high resolution feature maps generated by the CNN in the decoding path. Along with this, Chen et al. found that using a CNN and Transformer to encode features performed better than just using a Transformer as the encoder [3].

Patch Embedding. The feature maps from the CNN are tokenized by reshaping them into a sequence of 2D patches $\{x_p^i \in \mathbb{R}^{P^2 \cdot C} | i = 1, ..., N\}$, of size $P \times P$, where $N = \frac{HW}{P^2}$ is the number of patches. For our model, we will be using a patch size of 16×16. Using a trainable linear projection, the vectorized patches x_p are mapped into a latent D-dimensional embedding space. The equation that is used to learn the specific position embeddings that are to the patch embeddings is:

$$z_0 = [x_p^1 E; x_p^1 E; ...; x_p^N E] + E_{pos},$$

where the patch embedding projection is $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ and the position embedding is denoted by $E_{pos} \in \mathbb{R}^{N \times D}$.

Visual Transformer Encoder. The embedding sequence is passed through an encoder which 12 visual transformer layers. Each of the layers contain Multi-head Self-Attention (MSA) and Multi-Layer Perception (MLP) blocks, and the output of each layer ℓ can be written as:

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1},$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell,$$

where z_L is the encoded image representation and $LN(\cdot)$ is the layer normalization operator.

Cascaded Upsampler. The sequence of hidden feature $z_L \in \mathbb{R}^{\frac{HW}{P^2} \times D}$ from the encoder is reshaped to the shape of $\frac{H}{P} \times \frac{W}{P} \times D$, and passed through a 1×1 convolution to reduce the channel size of the reshaped feature to the number of classes. After reshaping, it is passed through four cascaded upsampling blocks, consisting of a $2\times$ upsampling operator, a 3×3 convolution layer, and a ReLU layer successively, in order to get from resolution $\frac{H}{P} \times \frac{W}{P}$ to $H \times W$ for predicting the final segmentation result.

The architecture used with the CNN, visual transformers, and cascaded upsampler allows using skip-connections for feature aggregation at different resolution levels.

3.2 Training Overview

Our model will be trained using 146 random samples (70%) from the KiTS19 dataset. The images will be resized to a resolution of 224×224 to fit the input layer of our model, and as mentioned in the architecture section, we will be using 16×16 for our patch size. The model will also be trained using a SGD optimizer, a learning rate of 0.01, a momentum of 0.9, weight decay of 1e-4, and a batch size of 24. These are the same hyperparameters the was used by Chen et al. [3]. One thing that we decided to change was reducing the number of training iterations from 14,000 iterations to 7,000 since our dataset had around double the number of samples as the Synapse dataset that they used.

4 Experimental Design

4.1 Main Purpose

The goal of this experiment is to determine how a successful medical image segmentation neural network architecture will perform in the new medical image segmentation task of renal tumor segmentation. Furthermore, our experiment will provide us with insights on how to improve the TransUNet architecture for better performance on KiTS19, and a deeper understanding of the characteristics of our data set [7].

4.2 Design

The data set will be divided into the same splits as what was used on TransUNet in multi-organ segmentation [3]. The 209 samples that make up KiTS19 will be split into a 70/10/20 train/validation/test split and a 70/30 train/test split. Using the Captum library, we will extract the information that our neural net learns. Based on these insights, we will determine what the neural network has

learned. By understanding what patterns in the data set the neural network was able to learn we can improve our understanding of the data set, and more importantly information about renal cancer and abdominal CT scans that were not apparent from human observation.

4.3 Evaluation Metrics

Integrated Gradients will show how invariant the network is to our dataset as well as how much certain features influence the prediction of the neural network. This will tell us the most relevant features of renal CT scans that predict cancer at a pixel level. The metric may also reveal a new way of reasoning about tumor segmentation that most human radiologists have not discovered. [10]. Furthermore Sundararajan et al. showed that integrated gradients by way of feature importance provided retina specialists with a level of trust in the neural network's predictions that they would not have had without the metric.

The second metric we will be using is the Dice Similarity Coefficient (DSC). The DSC measures the fraction number of pixels in the same location across two images that are exactly the same relative to the number of pixels in both images. The use of this metric is to measure how close our neural networks predictions are to the segmentation produced by a human radiologist, which is currently the gold standard in radiology [1].

References

1. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. Radiology **278**(2) (2016) 563–577 PMID: 26579733.
2. NIH: Financial burden of cancer care. (2021)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. (2021)
4. Barbera, G.L., Gori, P., Boussaid, H., Belucci, B., Delmonte, A., Goulin, J., Sarnacki, S., Rouet, L., Bloch, I.: Automatic size and pose homogenization with spatial transformer network to improve and accelerate pediatric segmentation. (2021) 1773–1776
5. Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M.: Vision transformers in medical computer vision – a contemplative retrospection (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020)
7. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
9. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. CoRR **abs/1506.02025** (2015)
10. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017)