ViT-UNet: Using visual transformers for renal tumor segmentation

Matt Nguyen, Neel Karsanbhai

University of Colorado, Boulder

1 Introduction

1.1 Motivation

The analysis of medical images (MRIs, CT scans, PET images) for cancer diagnosis, and cancer analysis on patients has is currently a high level analysis done by human radiologists [1]. This is inefficient from a cost perspective, and data extraction perspective. The radiomics is a practice based on the idea that medical imaging in general can provide more insights into a patient than just the presence of a tumor, or location of a tumor. 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.

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 [2]. 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 [3]. 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 [4].

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. [5], 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 [4]. Our proposed network will trained and tested using the KiTS19 dataset [6].

2 Related Work

2.1 Combining UNet with ViT for Renal Cancer Segmentation

- J. Chen et al. [2]
- Dosovitskiy et al. [5]
- Ronneberger et al. [7]

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.

2.2 UNet with Spatial Transformer Network for Renal Tumor Segmentation

- Berbera et al. [3]
- Jaderberg et al. [8]
- Ronneberger et al. [7]

While the network proposed by Berbera et al. was trained and tested on the same KiTS19 dataset [6] as we are, they used a spatial transformer network proposed by Jaderberg et al. [8] while we plan on using visual transformers.

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. 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)
3. 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
4. Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M.: Vision transformers in medical computer vision – a contemplative retrospection (2022)

5. 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)
6. 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)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
8. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. CoRR **abs/1506.02025** (2015)