UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Convolutional Networks versus Transformers: A Comparison in Prostate Segmentation.

Fernando Nicolás Vásconez González

Ingeniería en ciencias de la computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en ciencias de la computación

Quito, 3 de junio de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingeniería

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Convolutional Networks versus Transformers:
A Comparison in Prostate Segmentation.

Fernando Nicolás Vásconez González

Nombre del profesor, Título académico	**María Gabriela Baldeón, PhD**
Nombre del profesor, Título académico	**Daniel Riofrío, PhD**

Quito, 3 de junio de 2022

© DERECHOS DE AUTOR

Nombres y apellidos: Fernando Nicolás Vásconez González

Código: 00204187

Cédula de identidad: 1718586876

Lugar y fecha: Quito, 3 de junio de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El cáncer de próstata es uno de los tipos más comunes de cáncer que afectan a los hombres. Una manera de diagnosticar y tratar este tipo de cáncer es a través de manualmente segmentar la región prostática y analizar su tamaño o consistencia con imágenes de resonancia magnética. Sin embargo, este proceso requiere de un radiólogo experimentado, toma una buena cantidad de tiempo y es susceptible a errores humanos. Recientemente, las redes neurales convolucionales o CNN han sido utilizadas para automatizar la segmentación de la próstata. En particular, la arquitectura U-net se ha convertido en el estándar por su eficacia y rendimiento. No obstante, las CNN son incapaces de aprender sobre dependencias a un alcance elevado por lo que recientemente Transformers han sido traídos como una alternativa, obteniendo mejores resultados en el análisis de imágenes. A pesar de esto, los transformers han obtenido resultados competitivos cuando una gran cantidad de imágenes está disponible para el entrenamiento. En este trabajo se compararán dos arquitecturas U-net Residual y Unet TRansformers (UNETR) en la segmentación haciendo el grupo de imágenes ProstateX. Para analizar el efecto y el rendimiento de cada una de las arquitecturas este grupo de imágenes serán variadas en intervalos de 30 de las 120 imágenes en 3D. Los experimentos han demostrado que la arquitectura UNETR tiene un mayor rendimiento que U-net Residual. En promedio UNETR tiene una mejoría del 6% sobre el Unet Residual en base al Dice Score.

Palabras clave: Prostate Segmentation, Deep Learning, Transformers, Fully Convolutional Networks, Residual U-Net, UNetR.

ABSTRACT

Prostate cancer is one of the most common types of cancer that affects men. One way to diagnose and treat this type of cancer is by manually segmenting the prostate region and analyzing its size or consistency in MRI scans. However, this process requires an experienced radiologist, is time-consuming, and prone to human error. Recently, Convolutional Neural Networks (CNNs) have been applied to automate the segmentation of the prostate. In particular, the U-net architecture has become the de-facto standard given its performance and efficacy. However, CNNs are unable to model long-range dependencies. Hence, transformers have emerged as an alternative, obtaining better results than CNNs in image analysis. Nevertheless, transformers have obtained competitive results when a large dataset is available for training. In this work, a residual U-Net and the transformer UNetR are compared in the task of prostate segmentation using the ProstateX dataset. To analyze the effect the size of the dataset has on the performance, the training dataset is varied from 30 to 120 3D images. The experiments show that the transformer architecture has a better performance than the residual U-Net. In average, UNetR has an 6% performance increase in the test Dice Score over the residual U-Net.

Key words: Prostate Segmentation, Deep Learning, Transformers, Fully Convolutional Networks, Residual U-Net, UNetR.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCTION

Cancer is the second-leading cause of death around the world, affecting 1 of every 3 people. Cancer is a disease caused by the development of abnormal cells growing and dividing uncontrollably, infiltrating, and destroying the normal body tissue (Mayo Clinic Foundation, n.d.). Furthermore, it is an expensive disease that costs on average $123,400,000 annually per patient in medical services and medications (Yabroff, 2021). Prostate cancer is the second most frequent type of cancer in men (Rawla, 2019). This type of cancer is more likely to appear at older ages and is hard to detect because it has no symptoms until it is in advanced stages. Therefore, screening is usually recommended for men after turning 55 and at the start of any symptom (The American Cancer Society medical and editorial content team, 2019).

Many methods have been developed to screen for prostate cancer, such as prostate-specific antigen (PSA), ultra-sound guided, transrectal biopsy, and magnetic resonance imaging (MRI) (Eklund, et al., 2021). Although, there is no consensus on the test that should be applied to a patient (Eldred-Evans, et al., 2020), it is common to use the PSA or Directal Rectal Examination (DRE). However, both have their disadvantages. On one hand, PSA values could be affected by medications, medical procedures, prostate infection or enlarged prostate (Centers for Disease Control and Prevention, 2021). Meanwhile, DRE may result in a high number of false positives that could lead to unnecessary biopsy or over-diagnosis and over-treatment (Naji, et al., 2018). Prostate MRI analysis has gained popularity because it allows to identify areas of the prostate suggestive of cancer and improves the accuracy of cancer diagnosis (Eklund, et al., 2021). However, MRI analysis is time-consuming, subjective, and prone to human error. Moreover, the diagnosis given after analyzing the images may differ from expert to expert (Razzak, Naz, & Zaib, 2017). MRI was chosen over other type of screening images type because of the advantages that MRI provides like the increased soft

tissue contrast and better motion correction providing images with better resolution (Nie, Cao, Gao, Wang, & Shen, 2016).

Deep learning has improved the analysis of medical data by integrating enormous amounts of heterogeneous data for diagnosis and disease recognition (Lundervold & Lundervold, 2019). In the area of medical image analysis, Convolutional Neural Networks (CNNs) are the more popular architectures in deep learning due to their astonishing results on object recognition and segmentation (Yamashita, Nishio, Kinh Gian Do, & Togashi, 2018). CNNs extract features from data by applying convolution operations using a structure of artificial neurons which are intended to simulate a biological neuron to achieve the simulation of the visual perception (Li, Liu, Yang, Peng, & Zhou, 2021).

In the task of image segmentation, Fully Convolutional Networks (FCN) have become dominant. The FCN architecture consists of two symmetric paths, an encoder and a decoder. The encoder is a contracting path that extracts features of the images, and the decoder is an expanding path that extracts positions. Additionally, the encoder gradually down-samples the resolution of the images, getting the feature maps, to improve the computation and on the decoder upsamples and start learning via receptive fields (Wang, Li, Duan, & Shenghui, 2021). This is done through a nonlinear filter instead of nonlinear function used in a general deep net. Based on the FCN structure, various architectures have been derived for prostate segmentation such as the U-Net (Ronneberger, Fischer, & Brox, 2015), Z-Net (Zhang, Wu, Chen, Chen, & Tang, 2019), PSNet (Tian, Liu, Zhang, & Fei, 2018), 3D Chan-Vese (L. F. da Silva, et al., 2020), Residual U-Net (Kerfoot, et al., 2019), Densenet-like U-net (Aldoj, Biavati, Michallek, Stober, & Dewey, 2020), and Hybrid 3D-2D U-Net (Ushinsky, et al., 2021). With Densenet-like U-net, results of segmentation for the different parts of the prostate were 91.2% for the prostate 89.2% for central zone and 76.4% peripheral zone in dice score (Aldoj, Biavati, Michallek, Stober, & Dewey, 2020). Even though, CNNs have obtained an exceptional

performance, they struggle at capturing long-range information because of the regional locality of convolutional operations and its poor scaling properties (Ramachandran, et al., 2019). This work uses Residual U-net which is an enhanced model created over the U-net that improves the structure by adding residual connections on the encoding and decoding stages (Kerfoot, et al., 2019).

In natural language processing (NLP), transformers have become the algorithm of choice because of their computational efficiency and scalability. Due to their success in NLP applications, transformers have been implemented in image processing by splitting an image into patches and provided in sequence into the transformer (Dosovitskiy, et al., 2020). Moreover, transformers have come to improve the deep learning architectures, this is because of the innate global self-attention mechanisms (Jieneng, et al., 2021). Consisting solely of attention mechanism, stacks of self-attention, and pointwise fully connected layers, transformers are more parallelizable and require less time to train (Vaswani, et al., 2017). In computer vision, transformers overcome the lack of representation and process of high-level concepts in images which convolutions solely cannot solve (Wu, et al., 2020). Transformers architectures that have been developed for the task of medical image segmentation include the TrasU-Net (Jieneng, et al., 2021), TransBTSV2 (Li, et al., 2022), Swin UNETR (Hatamizadeh, et al., 2022), RTNet (Huang, Li, Xiao, Shen, & Xu, 2022), and U-netR (Hatamizadeh, et al., 2021). In particular, U-netR is used in this work as it has gained popularity in 3D Medical Image Segmentation. This network has achieved a 95% of Hausdorff distance in the segmentation of abdominal organs (Hatamizadeh, et al., 2021).

In computer vision, the main difference between CNN and Transformers is the way they analyze the image data. CNN learn by feature representations of the images this is done by convolution kernels that are used to compute the different feature maps (Gu, et al., 2018).While Transformers encode the images as a sequence of 1D patch embeddings and utilize self-

attention modules to learn (Hatamizadeh, et al., 2021). This allows transformers to learn the information of the image as it would learn in NLP.

Transformers have shown to outperform CNNs in computer vision tasks when large datasets are available. However, given their learning over-flexibility, transformers have a tendency of overfitting small datasets. Considering that in medical scenarios acquiring labelled datasets can be quite costly and time-consuming, it is indispensable to test their predictive performance in these applications. In this work, the transformer U-netR and the CNN Residual U-net are compared for the task of prostate MRI segmentation. Datasets ranging from 30 to 120 3D images from the PROSTATEx challenge are used to test the performance of the U-Net and UNetR architecture using the same test set. The results show that the transformer architecture performs better than residual U-Net even with fewer images which shows that transformers are capable of learning from context and high-level image features which is an improvement over CNN.

MATERIALS AND METHODS

Experimental dataset

The experiments are performed on a prostate MRI dataset from the PROSTATEx Challenge (2017)[1] (Radboud University Medical Centre, 2022). The dataset consists of volumetric MRI images from 150 patients. Images vary in sizes from (320×320×18) to (640×640×27), with an inter-slice resolution ranging from 0.3mm × 0.3mm to 0.6mm × 0.6mm, and intra-slice resolution between 3mm to 4.5mm. The data has been acquired from two different types of Siemens scanners: the MAGNETOM Trio and Skyra. The prostate gland, Central Zone, Transitional Zone, and Peripheral Zone have been annotated by expert radiologists of Moffit Cancer Center. Each image is read, transposed, and casted into 32-bit float. Normalization is applied using a pixel-wise linear transformation to a maximum value of 1 and the minimum value to 0, as shown in equation (1).

$$\text{outputPixel}=(\text{inputPixel}-\text{inputMin})\times\frac{(\text{outputMax}-\text{outputMin})}{\text{inputMax}-\text{inputMin}}+\text{outputMin} \quad (1)$$

Where input pixel is the pixel in a given position to be normalized, inputMin is the minimum pixel value in the image, inputMax is the maximum pixel value in the image and finally, the outputMax is 1 and outputMin is 0. To obtain a normalization between [0-1].

The images of the dataset are rescaled to size (256×256×32) and the voxel spacing of (0.5mm, 0.5mm, 1.5mm). Each voxel corresponds to a pixel and a slice thickness of the volumetric image. In order to resize the images and labels two interpolators are used with the help of Simpleitk library[2]. Specifically, the images are resampled with the b-spline method while the labels with the Nearest Neighbor technique. Also, the centroid is calculated for each image and label, to maintain the prostate in the center. The trim of the image may produce some information loss in each image, like the context around the prostate like the organs that

[1] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656
[2] https://simpleitk.org/

are surrounding the prostate. Furthermore, in case of enlargement of the prostate or its surroundings, it may lead to a false diagnosis. Last but not least, for training and testing the dataset was randomly divided using a 80:20 ratio for training and testing, respectively, using a 5-fold cross-validation. Each fold was stored as numpy arrays in different folders the same data at each fold is available for both models.

Models

The models compared are the Residual U-Net and U-NetR. The Residual U-Net is an encoder-decoder architecture as presented in Figure 1. The architecture has a total of 5 residual units in the encoder path and 4 up-sample units in the decoder path, resembling a U-shaped architecture. Each residual unit consists of a convolutional layer with a stride of 2 that down-samples the image to half, an instance normalization layer to prevent contrast shifting, a parametric rectifying linear unit (PReLU), a convolutional 2D, an instance normalization layer, and finally into a PReLU layer. Only the first residual unit has a stride of 1. Meanwhile, the up-sample units are composed of a transpose convolutional operation that doubles the size of the feature map, a convolutional layer, instance normalization layer, and PReLU activation function. The encoder and decoder paths are connected through a concatenation operation by the opposite residual and up-sample units. The benefit of these connections is that the low- and high-level details are considered to produce the final segmentation.

The second architecture implemented is the U-netR as shown in Figure 1. U-netR uses both a CNN and a Transformer structure. U-NetR is an architecture based on the U-Net model that implements a stack of transformers in the encoder path. The encoder is connected skiply with the decoder. Since transformers work on 1D input, the 3D images are transformed to 1D by flatenning them into uniform non-overlapping patches by $x_v \in R^{N\times(P^3C)}$ where (P,P,P) denotes the resolution of each patch and $N=(H\times W\times D)/P^3$ is the length of the sequence. Then a linear layer is used to project the patches into a K dimensional embedding space. This layer is constant

throughout the transformer layers. To preserve the spatial information of the extracted patches, an 1D learnable positional embedding is used $E_{pos} \in R^{N \times K}$ to the projected patch embedding $E \in R^{(P^3 C) \times K}$.Then a stack of transformer blocks is used, these blocks comprise multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers. These blocks consist of a normalization layer and in the MLP there are two linear layers with Gaussian Error Linear Unit (GELU) activation functions. In the MSA layers, there are parallel self-attention (SA) heads that calculates its weights by measuring the similarity between two objects according to the mapping of the representations finally a softmax function is used. Inspired by Unet, the features from multiple resolution of the encoder are merged with the decoder, then a sequence representation is extracted from the transformer and then reshaped as the input (Hatamizadeh, et al., 2021).

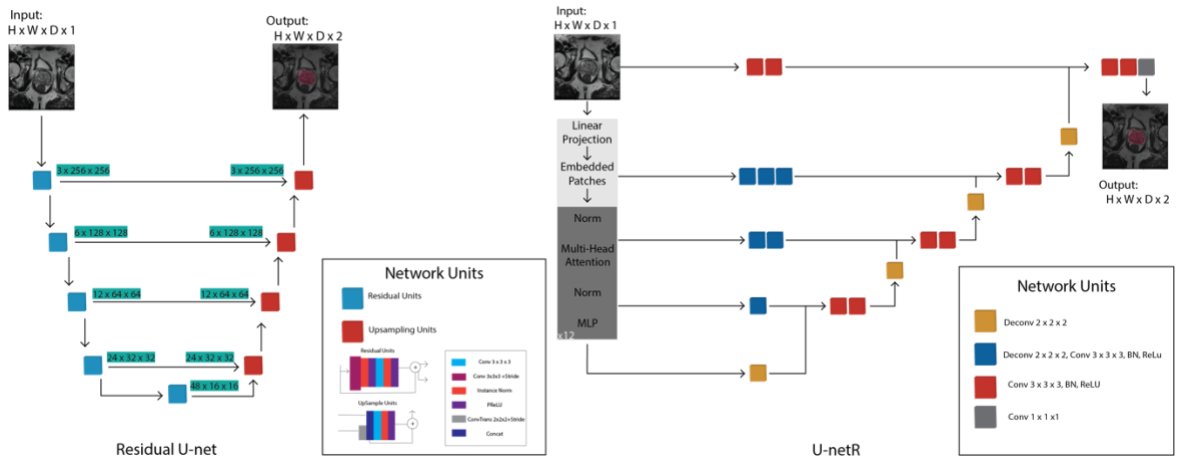

Figure 1: Residual U-net and UNETR architecture.

Experimental Setup

Training and Testing.

The architectures are implemented in PyTorch (PyTorch, 2022) and MONAI (Monai, 2022). Both models were trained using a RTX-3060 GPU, AdamW optimizer with a learning rate of 0.00001, and a batch size of 3 due to memory restriction. No pre-trained wieghts were used.

The size of the training set was varied during training from 30, 60, 90, and 120 images to evaluate the performance of each model as the dataset increases were used and varied in both models to test the performance of each. The loss function optimized during training is a combination of the soft dice loss and cross-entropy loss as displayed in Eq.(2). In which I is the number of voxels, J is the number of classes, $Y_{i,j}$ is the probability output and $G_{i,j}$ is the ground truth, for class j at voxel i. The test dataset was not changed in any fold, being the same test dataset for every training set in order to have a fair comparison. In an interval of 5 epoch the model is tested using the Dice Score Eq.(**3**), 95% Hausdorff Distance (HD) Eq.(**4**) and Jaccard Distance Eq.(**5**). The 95-percentile Hausdorff Distance is a distance metric that calculates the maximum distance between the ground truth and the nearest point of the segmented zone. The 95th percent of the boundaries are reported to eliminate the impact of outliers. The Dice Score and Jaccard Distance are overlap based measures. The Dice measures the volumetric overlap between the predicted segmentation and the ground truth segmentation, while the Jaccard Distance calculates the extent of overlap between the ground truth and the prediction zone.

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j} \quad (2)$$

$$\boldsymbol{Dice(G,P)} = \frac{\boldsymbol{2\sum_{i=1}^{I} G_i P_i}}{\boldsymbol{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}} \quad (\mathbf{3})$$

$$\boldsymbol{HD(G',P') = max\{max_{g'\in G'} min_{p'\in P'} \left||g'-p'|\right|, max_{p'\in P'} min_{g'\in G'} \left||p'-g'|\right|\}.} \quad (\mathbf{4})$$

$$\boldsymbol{\mathcal{D}_{\mathcal{J}}(G',P') = \frac{|G'\cup P'| - \sum_{i=1}^{I} G_i' P_i'}{|G'\cup P'|}} \quad (\mathbf{5})$$

Selection Criteria.

The metrics to determine the best model are the Dice and the 95 Hausdorff distance, the other metrics shown are to check congruence and the performance of each prediction in

comparison with the ground truth. The results are an average all 5-folds with its respective standard deviation. The best loss values are when they are low, dice metrics are calculated over percentage therefore higher values are better. The other metrics, since they are distances, lower values are better because it shows that the ground truth and the label predicted are closer, i.e. a better segmentation work.

RESULTS AND DISCUSSION

The experiments done in all the groups of images of the dataset demonstrate improved performance of UnetR over Residual U-net. Specifically, UnetR achieves a better segmentation dice score even when the data is scarce. When the data is scarce UnetR shows a better performance of a 6% over Residual Unet as seen in Table 1. This is because UnetR uses transformers and also because of the attention mechanisms it captures both global and local dependencies improving segmentation. In an average UnetR outperforms 6% over CNN. In Figure 2 the predictions made by the models show that when trained on scarce data on UnetR the prediction is closer and have less variations as the prediction of Residual Unet. Furthermore, when trained on the complete dataset the predictions are more accurate and the borders are closer to the expected label.

Moreover, as shown in the Table 1these results are congruent with the loss score, showing that when you have a larger dataset the models reduce the information lost. Furthermore, as seen on the segmentation results Figure 2, while the dataset is larger the segmentation improves, when the dataset is of 30 the predictions are more misshapen, and the borders are misplaced but with the augmentation of the data the predictions start to be more accurate and the borders became closer to the ground truth. The best model was the UnetR with a dice score of 0.84. Nevertheless, Jaccard score of this model shows that the prediction even when is closer to the ground truth has more difference than the prediction of the residual Unet. Finally, the

graphs of the Loss versus Epochs for each group of data were analyzed in order to check if there was an overfitting on the models and to ensure the correct learning as seen on Figure 3. Both models show a fair learning and no overfitting.

Arch.	UNETR			
Data	Loss ± σ	Dice ± σ	Jaccard ± σ	95 HD ± σ
120	0.32 ± 0.14	0.84 ± 0.03	0.73 ± 0.04	9.90 ± 3.90
90	0.46 ± 0.12	0.83 ± 0.04	0.71 ± 0.05	11.85 ± 4.68
60	0.45 ± 0.12	0.82 ± 0.04	0.70 ± 0.05	12.93 ± 4.14
30	0.76 ± 0.05	0.74 ± 0.11	0.61 ± 0.11	34.52 ± 22.04
Arch.	Res. U-net			
Data	Loss ± σ	Dice ± σ	Jaccard ± σ	95 HD ± σ
120	0.56 ± 0.16	0.75 ± 0.06	0.61 ± 0.07	20.82 ± 6.59
90	0.61 ± 0.13	0.74 ± 0.04	0.60 ± 0.05	23.46 ± 5.34
60	0.69 ± 0.07	0.71 ± 0.06	0.57 ± 0.06	29.92 ± 10.30
30	0.76 ± 0.05	0.61 ± 0.11	0.61 ± 0.11	34.52 ± 22.04

Table 1: Average Results obtained from UnetR and Residual Unet for the different datasets groups

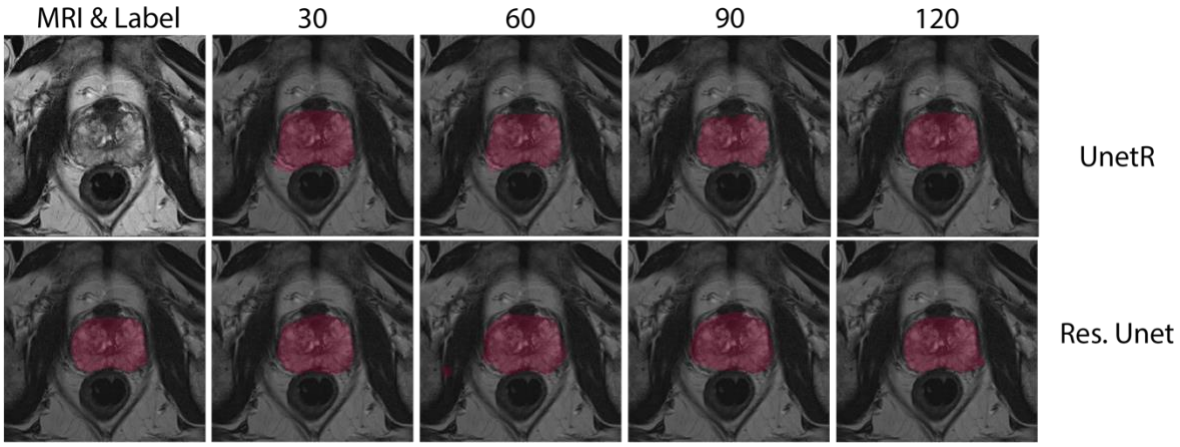

Figure 2: Results of UnetR and Residual Unet segmentation, on the first row the predictions of UnetR. On the second row the predictions of Residual Unet.

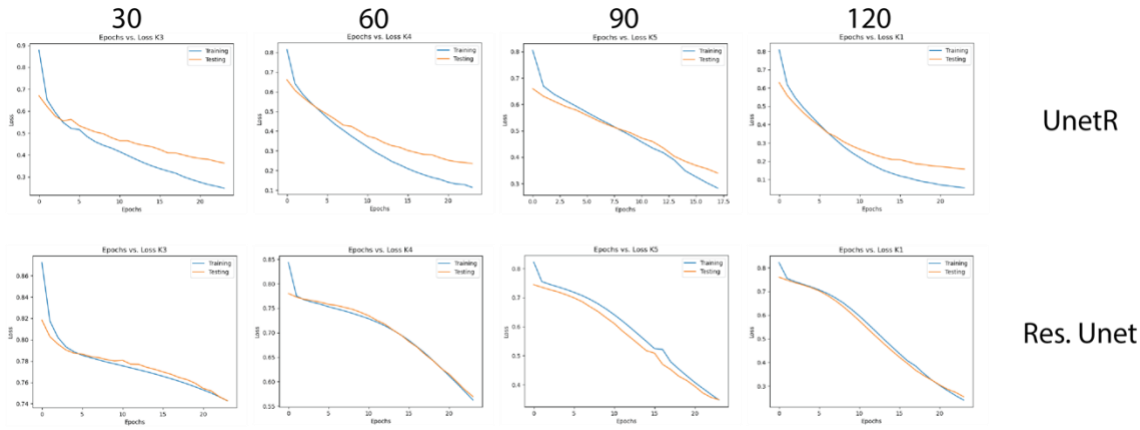

Figure 3:Training plots comparison (UnetR vs. Residual Unet) - Epochs vs. Loss: No signs of overfitting is shown.

CONCLUSIONS AND FUTURE WORK

Transformers have come to improve computer vision with deep learning. The experiments have shown that convolutional layers combined with transformers have a great improvement in segmentation. In this case study, prostate segmentation, transformer networks improve 6\% over CNN networks. When the data is scarce, both models have similar performance but as the data increases so the performance of the Transformers network. For further investigations Transformers could improve even more with more data or using data augmentation. Finally, using Transformers and CNN networks could be helpful in the segmentation of prostate zones and cancer classification.

ACKNOWLEDGMENT

Authors would like to thank Moffit Cancer Center and the Applied Signal Processing and Machine Learning Research Group of USFQ for providing the database and the computing infrastructure (NVidia DGX workstation) to implement and execute the developed source code, respectively.

REFERENCIAS BIBLIOGRAFICAS

Aldoj, N., Biavati, F., Michallek, F., Stober, S., & Dewey, M. (2020). Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Scientific Reports*.

Centers for Disease Control and Prevention. (2021, August). *What is screening for prostate cancer?* Retrieved from https://www.cdc.gov/cancer/prostate/basic_info/screening.htm

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*.

Eklund, M., Jaderling, F., Discacciati, A., Bergman, M., Annerstedt, M., Aly, M., . . . Nordstrom, T. (2021). MRI-Targeted or Standard Biopsy in Prostate Cancer Screening. *New England Journal of Medicine*, 908-920.

Eldred-Evans, D., Tam, H., Sokhi, H., R. Padhani, A., Winkler, M., & Ahmed, H. (2020). Rethinking prostate cancer screening: could MRI be an alternative screening test? *Nature Reviews Urology*, 526-539.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., . . . Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 354-377.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images.* Retrieved from Arxiv: https://arxiv.org/abs/2201.01266

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., . . . Xu, D. (2021). *UNETR: Transformers for 3D Medical Image Segmentation.* Retrieved from Arxiv: https://arxiv.org/abs/2103.10504

Huang, S., Li, J., Xiao, Y., Shen, N., & Xu, T. (2022). RTNet: Relation Transformer Network for Diabetic Retinopathy Multi-lesion Segmentation. *IEEE Transactions on Medical Imaging*.

Jieneng, C., Yongyi, L., Qihang, Y., Xiangde, L., Ehsan, A., Wang, Y., . . . Zhou, Y. (2021). *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.* Retrieved from Arxiv: https://arxiv.org/abs/2102.04306

Kerfoot, E., Clough, J., Oksuz, I., Lee, J., P. King, A., & A. Schnabel, J. (2019). Left-Ventricle Quantification Using Residual U-Net. *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, 371-380.

L. F. da Silva, G., S. Diniz, P., Ferreira, J. L., V. F. Francca, J., C. Silva, A., C. de Paiva, A., & A. de Cavalcanti, E. (2020). Superpixel-based deep convolutional neural networks and active contour model for automatic prostate segmentation on 3D MRI scans. *Medical & Biological Engineering & Computing*, 1947-1964.

Li, J., Wang, W., Chen, C., Zhang, T., Zha, S., Wang, J., & Yu, H. (2022). *TransBTSV2: Towards Better and More Efficient Volumetric Segmentation of Medical Images.* Retrieved from Arxiv: https://arxiv.org/abs/2201.12785

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift fur Medizinische Physik*, 102-127.

Mayo Clinic Foundation. (n.d.). *Cancer-Symptoms and causes*. Retrieved Mayo 14, 2022, from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/cancer/symptoms-causes/syc-20370588

Monai. (2022). *MONAI Home*. Retrieved May 13, 2022, from Monai: https://monai.io

Naji, L., Randhawa, H., Sohani, Z., Dennis, B., Lautenbach, D., Kavanagh, O., . . . Profetto, J. (2018, Marz). Digital Rectal Examination for Prostate Cancer Screening in Primary Care: A Systematic Review and Meta-Analysis. *The Annals of Family Medicine*, 149-154. Retrieved from https://doi.org/10.1370%2Fafm.2205

Nie, D., Cao, X., Gao, Y., Wang, L., & Shen, D. (2016). Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks. *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 170-178.

PyTorch. (2022). *Install PyTorch*. Retrieved May 13, 2022, from PyTorch: https://pytorch.org

Radboud University Medical Centre. (2022, May 07). *PROSTATEx - Grand Challenge*. Retrieved from https://prostatex.grand-challenge.org

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-Alone Self-Attention in Vision Models. *CoRR*.

Rawla, P. (2019). Epidemiology of Prostate Cancer. *World Journal of Oncology*, 63-89.

Razzak, M. I., Naz, S., & Zaib, A. (2017, April 22). *Deep Learning for Medical Image Processing: Overview, Challenges and Future.* Retrieved from Arxiv: https://arxiv.org/abs/1704.06825

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*.

SimpleITK Library. (2022, May 16). Retrieved from SimpleITK Library: https://simpleitk.org/

The American Cancer Society medical and editorial content team. (2019). *Can prostate cancer be found early?* Retrieved from https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/detection.html

Tian, Z., Liu, L., Zhang, Z., & Fei, B. (2018). PSNet: prostate segmentation on MRI based on a convolutional neural network. *Journal of Medical Imaging*, 1-6.

Ushinsky, A., Bardis, M., Glavis-Bloom, J., Uchio, E., Chantaduly, C., Nguyentat, M., . . . Houshyar, R. (2021). A 3D-2D Hybrid U-Net Convolutional Neural Network Approach to Prostate Organ Segmentation of Multiparametric MRI. *American Journal of Roentgenology*, 111-116.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *CoRR*.

Wang, L., Li, R., Duan, C., & Shenghui, F. (2021). Transformer Meets DCFAM: A Novel Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *CoRR*.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Tomizuka, M., . . . Vajda, P. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *CoRR*.

Yabroff, K. R. (2021). Annual Report to the Nation on the Status of Cancer, Part 2: Patient Economic Burden Associated With Cancer Care. *JNCI: Journal of the National Cancer Institute*, 1670-1682.

Yamashita, R., Nishio, M., Kinh Gian Do, R., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 611-629.

Zhang, Y., Wu, J., Chen, W., Chen, Y., & Tang, X. (2019). Prostate Segmentation Using Z-Net. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*.