Kidney Semantic Segmentation using U-Net

William Han
University of California, Irvine

May 14, 2022

1 Introduction

Kidney cancer is among the most common pathologies in adults internationally and its prevalence is known to be increasing [**1**]. However, through advanced medical technology, kidney tumors are often discovered early while they are still manageable [**1**]. Kidney tumors are known for their obvious appearances in computed tomography (CT) imaging, which allow expert radiologist to deeply study different characteristics of it [**1**]. For further research, the recent and unexpected boom in convolutional neural networks (CNN) has been a great aid in streamlining detection of not only kidney tumors, but other pathological modalities. CNNs are widely used in medical computer vision, and more specifically, semantic segmentation, a pixel by pixel classification task utilized for highlighting regions of interest. In this study, I will be attempting to re-create 3 different versions of a very popular architecture called U-Net [**2**] for kidney tumor segmentation.

2 Methods and Materials

The data used in this project consists of kidney tumor CT exams extracted from the Kidney Tumor Segmentation Challenge (KiTS) [**1**]. 321 images and its respective mask were allocated to the training set, while 81 were used for validation. I created a 2D U-Net, 3D U-Net, and custom, Squeeze and Excite [**3**] U-Net architecture to accomplish the segmentation task.

The 2D U-Net has 56 layers, which is including the input, normalization, regularization, concatenation, transpose layers. It also has 230,498 trainable

and 1,072 non-trainable parameters. Although the images passed into the 2D U-Net are 2 dimensional, I used 3D convolutional operations due to the fact that the full matrix was still a 3D tensor, where in (z, y, x), z is equal to 1. The input shape of the image is a 1 x 96 x 96 x 1 matrix and the kernel size used throughout was 1 x 3 x 3.

The 3D U-Net has 57 layers, 689,140 trainable parameters and 1,072 non-trainable parameters. In this network, images of size 96 x 96 x 96 were passed, therefore 3D convolutional operations were also used. Due to the different dimensionality, the kernel's size was 3 x 3 x 3 and the he normal kernel initializer was used to instantiate the weights of the model. Deep supervision was established by generating a series of predictions (logits) during different resolutions by subsampling. I applied this technique on the 9th and 8th layers of the expanding section of my model, resulting in logits outputted by subsampling of the original resolution (96) and half the original resolution (48) respectively. To reflect the different logit scores acquired through deep supervision, the MaxPool operation for every logit ensured that as my predictions decreased, the most important target is not lost.

In the custom, Squeeze and Excite U-Net, there were 76 layers, 234,580 trainable parameters, and 1,072 non-trainable parameters. The images passed into this model were the same as the 2D model mentioned previously, thus 1 x 96 x 96 x 1 in size. The kernel size (1 x 3 x 3) and padding sequence (same) were also alike the 2D model. Squeeze and Excitation operations were applied after every contracting layer.

The batch sizes for the 2D U-Net and Squeeze and Excite U-Net were 16. Initially, the batch size for the 3D U-Net architecture was 2, but I decided to change it to 16 for consistency. A same padding sequence was initialized to conserve the spatial dimensions of the inputs and outputs for each layer. The optimizer used was Adam and the learning rate stayed at 2e-4 throughout. A sparse categorical cross entropy loss function was used during training and for interpretation, the dice score metric was applied. After training all 3 models for 10 epochs, I planned to evaluate the model by collecting the mean, median, 25th percentile, and 75th percentile for the test sets during training and validation.

3 Results

I evaluated the model by creating a NumPy array of dice scores for the test set during training and validation. I distributed these dice scores into a Pandas data frame. With these dice scores and through Pandas operations, I found the 25th and 75th percentile, as well as the mean and median. It should be known that these scores were rounded to 3 decimal points. These statistical measures gave me an accurate performance measure of each of my models. The highest dice score for each model in its respective statistical measure category is denoted in bold.

Models	Mean	Median	25th percentile	75th percentile
2D	0.954	0.964	0.950	0.973
3D	**0.964**	**0.970**	**0.961**	**0.975**
Custom	0.950	0.964	0.946	0.971

Table 1: Statistics for Training Test Set

Models	Mean	Median	25th percentile	75th percentile
2D	0.923	**0.956**	0.912	**0.969**
3D	**0.933**	**0.956**	**0.936**	0.967
Custom	0.931	**0.956**	0.926	**0.969**

Table 2: Statistics for Validation Test Set

Note that during validation, the median for all three models were synonymous. Additionally, the 75th percentile for the 2D U-Net model and the custom U-Net model were the same. Overall, it seems that that 3D U-Net model generally outperforms the others.

4 Discussion

The results were not expected because I hypothesized that the custom model would be best in performance. The biggest difference between the three different algorithms is the input dimensionality. The Squeeze and Excite model was given 2 dimensional data (1 x 96 x 96), therefore we cannot say how well it would perform on 96 x 96 x 96 images. I chose the Squeeze

and Excite model due to its scalability in terms of feature maps. It would be interesting to experiment on 3D images for the custom model for future work.

References

[1] "The 2021 Kidney Tumor Segmentation Challenge." *Kits21*, kits21.kits-challenge.org. Accessed 5 May 2022.

[2] "U-Net." *Wikipedia*, 3 Aug. 2020, en.wikipedia.org/wiki/U-Net. Accessed 5 May 2022.

[3] Hu, Jie, et al. "Squeeze-And-Excitation Networks." ArXiv:1709.01507, 16 May 2019, arxiv.org/abs/1709.01507. Accessed 5 May 2022.