

Comparison of Classification and Segmentation Models in Kidney Tumor Detection

William Han
University of California, Irvine

June 9, 2022

1 Introduction

Kidney cancer is a common world-wide disease with an increasing prevalence [1]. Therefore, medical practitioners and researchers have dedicated countless hours in developing next generation technologies that will try to combat this pathology. In parallel, the field of machine learning has seen much success in image classification and segmentation using convolutional neural networks (CNN). The prognosis of kidney cancer is mainly determined by how early one detects it. Computed tomography (CT) imaging is a popular method radiologists use to uncover richer characteristics during deep inspection for kidney tumors [1]. The collaboration between medical experts and computer scientists has made it possible to use CNNs for classifying and segmenting tumors in CT images. In this study, I will build 2 models for classification and a single model for segmentation on 2D CT scans of the kidney. The first classification model is a basic CNN while the second classification model uses a CNN with a Squeeze and Excite architecture. The segmentation model replicates the U-Net architecture [2]. I will then proceed in a comparison study by computing metrics for all three models and acknowledging important differences between them in regards to performance and architecture.

2 Methods and Materials

The data used in this project comes from the Kidney Tumor Segmentation Challenge (KiTS) [1]. It consists of 402 images and masks of kidney tumor CT exams. 321 images and masks were allocated to the training set and 81 for the validation set.

Each 2D CT scan had a dimensionality of $1 \times 96 \times 96 \times 1$ and a kernel size of $1 \times 3 \times 3$ was used consistently. I used the 'same' padding method and 'he normal' as the kernel initializer. The sparse categorical cross entropy loss function was used throughout. In regards to metrics, I computed the accuracy, specificity, positive predictive value (PPV), and negative predictive value (NPV) for training test set and validation test set. A batch normalization and ReLu activation function was applied after every convolutional layer.

For the first classification model, there were 36 layers, which includes the input, normalization, regularization, and convolutional layers. There were 223,986 total parameters, with 223,010 trainable and 976 non-trainable. The second segmentation U-Net model consists of 56 layers, 231,570 total parameters (230,498 trainable and 1,072 non-trainable). The third classification model utilizes the Squeeze and Excite technique after every other contracting layer and has a total of 48 layers. It has a total of 228,646 parameters, with 227,670 being trainable and 976 non-trainable.

Due to the nature of the dimensionality (2D) for each CT scan, I had to implement a reduction strategy to fuse multiple predictions into a uniform prediction for each volume. For each prediction, I aggregated the exam score by adding all of the binarized validation predictions together. Taking these reduced scores, I applied a mean threshold to calculate the accuracy, sensitivity, PPV, and NPV.

3 Results

Overall, the 2D custom classification model performed most optimal with a validation accuracy of 0.6420. The segmentation U-Net model had the worst validation accuracy (0.5802). Note that during training, the specificity and positive predictive value for all three models are synonymous.

Models	Train Acc.	Train Spec.	Train PPV	Train NPV	
2D Classification	0.9315	0.8616	1.0000	1.0000	0.8804
2D Segmentation	0.7383	0.4717	1.0000	1.0000	0.6585
2D Custom Classification	0.9502	0.8994	1.0000	1.0000	0.9101

Table 1: Statistics for Training Test Set

Models	Valid Acc.	Valid Spec.	Valid PPV	Valid NPV	
2D Classification	0.6296	0.4681	0.8529	0.8148	0.5370
2D Segmentation	0.5802	0.3191	0.9412	0.8824	0.5000
2D Custom Classification	0.6420	0.4681	0.8824	0.8462	0.5455

Table 2: Statistics for Validation Test Set

4 Discussion

The results were expected because I hypothesized that the custom, Squeeze and Excite classification model would be the best in performance. It is difficult to compare the segmentation model with the two classification models because of the different assigned tasks. However, I speculate that the U-Net segmentation model and the first classification model performed the worst because of the architecture’s simplicity. Applying more convolutional layers and implementing residual connections or even the Squeeze and Excite method may boost their performance. For future works, I would like to expand the first classification model and the segmentation model by adding more layers and apply augmentation to the data.

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.