Politecnico di Torino

Course of Elaborazione Di Immagini Mediche

Technical report:

Prostate gland segmentation in T2 weighted MRI images with Unet

Donadio Serena, 279917

Facchi Ilaria, 279892

Parisi Giuseppe, 280062

Pasqua Burçak, 281935

1. Introduction

Prostate gland segmentation is still an open problem, which start with the volume extraction from MR images and leads to the choice of a treatment in case of pathology or prevention. Above all, mortality can be reduced thanks to an accurate early diagnosis.
Nowadays the gold standard for the prostate gland segmentation from magnetic resonance images is still the manual one realized by an expert operator: manual segmentation is, indeed, still the most precise technique despite it has some drawbacks such as inter and intra-operator variability, low reproducibility and it is very time consuming.
For these reasons, it is necessary to find an automated, operator-independent methodology.

In the last years, deep learning algorithms have evolved and showed interesting results in medical imaging, included segmentation problems, but even for algorithms segmentation is not an easy task for many reasons:

- In MR images prostate gland appears with no well-defined boundaries, and this may cause a complex differentiation from the tissues surrounding it, leading to under or over-segmentation.
- Different MR techniques or even different machines for a specific technique lead to large variations in signal intensity and noise on image.
- The size of the prostate gland varies in a wide range, and the same occurs for the shape.
- The inter-patients and inter-tissues variability in case of pathologies also add complexity to the problem.

In this report we will discuss on the application of a U-NET architecture model to perform prostate gland segmentation given a dataset of 50 T2-weighted MRI volumes. For training set and validation has been provided manual masks segmented by an experienced worker.

2. Methods

2.1 Deep CNN Architectures for Prostate Segmentation

The U-Net architecture was selected. This is made up of two convolutional pathways one performing downsampling and the other performing upsampling. Localization is accomplished on the upsampling pathway with skip connections, while classification is executed on the downsampling pathway.

The problem of segmentation, in this case is reduced to a binary classification problem. Each pixel is assigned to the prostate class or the background class. Taking different tests, was found that the U-net model densenet201 gave us better performances.

2.2. MRI Dataset

Our dataset was composed of T2-weighted MR volumes, each of dimension 512×512×24 (width x height x slices).

During training data has been fed to our model one slice of RM a time:

- 768 MR images for the training dataset (32 volumes);
- 192 MR images for the validation dataset (8 volumes);
- 240 MR images for the testing dataset (10 volumes);

Manually segmented masks by expert radiologist (gold standard) were used as ground truth for network training and validation, while the manual masks of the test set are held out by the organizer for further independent testing of the developed CAD system.

2.3. Image Pre-Processing

Taking into account that in the original images the background pixels were dominant with respect to the ones of the prostate, this leading to binary cross-entropy and binary accuracy biased toward background class, we first

cropped the images to 256x256 size. The crop has been centered on a visually evaluated region containing the prostates in our volumes.

The reduction of background pixels reduced the computation time and improved the results of the prostate segmentation in terms of DSC score (more balanced precision and recall).

Further we reduced the noise using a median filter correlating a kernel [3X3] on each RM slice.

2.4. Data Augmentation

To improve the accuracy of the training network and to avoid overfitting, data augmentation on the training dataset was applicated. Data augmentation essentially increases the spatial variation of the images and improves performance of models trained on small datasets.

To enlarge the training dataset random rotation in range (-5°,5°) and random zoom out in range (0,-50%) were used.

Our first models showed a poor accuracy in the segmentation of the firsts e the lasts slices of prostate, so it has been decided to use also an additional static data augmentation on the border slice of each 3D volume (factor 4 augmentation has been executed) to improve performances of the system on these slices.

2.5 . Training

Training parameters was set as follow:

- Maximum number of epochs of training: 20
- Batch size: 2
- Optimizer: Adam with an initial learning rate of 1e-3
- Loss function: binary cross-entropy
- Metrics: binary accuracy

In order to obtain a faster and more accurate model's convergence, we used pre-trained weights of the ImageNet database.

During training an early stopping criterion has been used. Due to this training sessions never went over 10 epochs and the limit of 20 epochs has never been reached. The early stopping criterion we used stopped the training if the model did not show improvement on the validation set in terms of binary accuracy over the limit of 1e-3 in 3 epochs. A little step of 1e-3 has been preferred to reach a finer convergence, but an adequate patience is needed to not prematurely stop (we set this to 3 epochs).

To obtain better results and test the robustness of preprocessing and training pipeline, a multi-start optimization approach has been used.

Further we decided to use our 3 best obtained models in terms of DSC score to build an ensemble CAD system.

2.6.Post processing

We observed that the ensembled model sometime gave masks with little imperfection such as little holes in the predicted masks and little masked objects around, so we performed small object deleting and small holes filling with a closure of kernel size 5x5. Small object maximal removal size was set to 1600 (area in pixel) which has been detected as a good limit for our case.

We observed that the system still did not perform well on slices at the border of the prostate gland. In order to try to partially avoid this, after observing that the system was predicting false areas with some holes at the border of the predicted prostate we checked this condition deleting these areas.

2.7. Evaluation Metric

We evaluated the performances of the CAD system for the segmentation of the prostate in T2W MRI the using dice similarity coefficient (DSC) calculated on each volume (we did not calculated DSC on each slice because this would not be sensitive on the entity of the fails in case true positives (TP) on the slice are 0). DSC values range between 0 to 1, with 1 indicating the best-case scenario.

3. Results and Discussion

3.1.Performance

On training set and validation set we obtained the following performances using as gold standard the manually segmented masks:

Training dataset		
Precision	Recall	F-score
0.9443	0.9709	0.9573

Validation dataset		
Precision	Recall	F-score
0.9138	0.9366	0.9240

3.2.Discussion

In our case we applied a little data augmentation because we observed that a more significant one decreases the performance and was not useful for our case in which images fed to the system were structured all the same (rectum on the right, prostate gland at the center and abdomen on the left). Due to this, performances on general images (structures not conform to the previously described one) were significantly worse (DSC on training 0.5870 and on validation 0.5635 has been obtained feeding the network with the very same images but rotated of a randomly chosen angle in the set 0,+90,+180,+270).
We also observed that despite the quite complex post-processing we performed only a very slight improvement has been obtained on DSC score on validation set and training set (0.1145 on training 0.0174 on validation).

4. Conclusions

Our model is quite computationally heavy but our only goal was to improve at best the performances in term of DSC score. Very similar but slightly poorer results could be obtained using only one model without ensembling three trained models and also skipping the post-processing pipeline used.

So we can conclude that our system probably is not the best in term of computation/result but is the best we obtained in terms of only performances. Also the system do not perform very well on the border slices of the prostate.