Medical ultrasound classification and segmentation

Daniel Santosh

1 Introduction

Real-time segmenting of the prostate gland from the transrectal ultrasound images is an exciting prospect for urology. Such a tool can aid in localising regions of interest during various urologic procedures, such as ablation therapy and needle biopsies. However, the boundaries of the prostate gland capsules is often not always clear in images, and is a challenge even for experienced urologists to identify. As such, this report presents a 2D image segmentation tool based on a UNet encoder-decoder network, whilst also utilising the VGG16 classification model to separate images with and without prostate glands. The objective is to develop and compare different model development strategies with available data set and labels from multiple observers, in the hopes of developing a real-time model for future applications in urologic surgeries.

2 Image Segmentation - Method

2.1 Sampling

Labelled transrectal ultrasound images were acquired for 200 cases, each consisting of numerous frames of size 58×52. Each frame was annotated with three segmentation labels from three independent observers, indicating the region of the prostate gland as binary masks. During training, at each epoch, all 200 cases were sampled without replacement, ensuring equal occurrence between cases. For each case, one of the numerous frames were randomly sampled, ensuring that all frames were equally likely to be sampled. In addition, different cases had different number of images, resulting in various sampling frequencies for different cases. The corresponding segmentation labels for each frame were sampled in two methods. For the first segmentation label sampling method, one of the three of the labels was randomly assigned to the frame. For the second segmentation level sampling method, a consensus label was assigned to frame. The consensus label was determined by a majority vote at the pixel level between the segmentation labels. This was achieved by summing the binary labels of the frame. For each pixel in the summed label, the pixel value was set to 1 if the summed value was above 1.5, or 0 otherwise. These two sampling methods were implemented using Pytorch datasets, which stored the frames and the corresponding segmentation labels for every case in our data sample. These two segmentation label sampling methods were implemented to gain insights into which sampling method would result better model accuracy after training.

2.2 Network Architecture

For the image segmentation of the prostate gland, an encoder-decoder like neural network was built, inspired by the UNet [1] and ResNet [2] architectures. Similar to the UNet architecture, the model implemented a contracting path followed by an expansive path. Along these paths, a repeating ResNet, consisting of residual blocks, was implemented after every down-sampling and up-sampling step, replacing the double convolutions in the original UNet. Residual blocks were preferred as they increased the depth of the convolutional neural network without losing information. Each residual block applies two 2×2 convolutions, each followed by a 2-D batch normalisation, a rectified linear unit (ReLU) and a dropout layer. At the end of the residual block, a residual shortcut connection was added, where the initial image is added to the final image. Furthermore, the number of features was expanded by a factor of 4 in the first convolution before a condensing of the same factor in the second convolution (ensuring the correct dimensions for the residual shortcuts). The ResNet itself contains one convolution, where the number of features is increased by a factor of 2 (apart from first and last layers of the network, where convolutions are used to map the 64 feature vector to the correct dimensions). This is followed by a 2-D batch normalisation, a ReLU, two residual blocks and a 2×2 max pooling with a stride of 2 for down-sampling. For the expansive path of the UNet, the max pooling is a replaced by a 2×2 up-convolution. Before applying the ResNet after each up-sample, skip layers are utilised to concatenate the corresponding cropped feature map from the contracting path as outlined in the original UNet. This network expands the feature space of a 1 channel input image from 32 up to 1024, before diminishing it back during the expansive path. On the other hand, the image size acts in the opposite manner, as it shrinks and expands as it passes through the network.

2.3 Training

The network was trained twice for the two segmentation labels methods, using the Adam optimization algorithm to optimise the 2-D dice loss function. The Adam optimizer was used as it combines properties of other effective stochastic gradient descents such as AdaGrad and RMSProp, leading to good results for spare gradients and noisy data [4]. For both models, bootstrap aggregating was used as an ensemble method to regularise our network, in the hopes of building multiple weak learners which will result in a higher accuracy compared to a single strong learner. In addition, the variance is reduced, hence preventing the overfitting of models during training. For both datasets (with different label sampling methods), the data was split in half, and two separate models were trained on the split data. For each sample split, a further split was carried out to obtain training and validation data for training the models. Data augmentation was carried out on the training data, to ensure robustness given the sparse dataset. Examining the ultrasound images, it is evident that there is no clear orientation or size to the prostate gland. Therefore, the training data was augmented by applying geometrical transformations, such as rotations, horizontal and vertical flips. In turn, this allowed our models to learn more orientations of the prostate gland, ensuring a better segmentation. The augmented training data and the validation data were randomly sampled using a data loader in mini-batches of size 8. Then, the segmentation models were trained for 500 epochs. During each epoch, the training data loaders was passed into the model and the 2-D loss between the predictions and the labels was calculated and optimized through backpropagation. The average training 2-D dice loss and accuracy, using the IoU loss (validation metric), were outputted after each epoch. At each epoch, after training, the average loss and accuracy of the model on the validation data set is outputted, to identify an potential over fitting. After training, predictions made on a holdout test data from both seperately trained models were averaged using a consensus vote.

3 Image Classification - Method

Similar to the segmentation datasets, classification datasets were constructed to contain frames and their corresponding classification labels. These labels indicated the presence of a prostate in an image frame. Binary classification labels for each image frame were determined by a majority vote between the three labels at image level. In other words, if a frame had two or more non-zero segmentation labels, it would indicate the presence of a prostate gland. Otherwise, the frame was assumed to not have a prostate present. This was implemented by counting the number of non-zero label tensors, and depending on whether the count was above or below 2, the frames were given either a binary vector classification of [1,0] (contains prostate) or [0,1] (does not contain prostate).

To the classify the ultrasound images, the VGG16 model [3], a convolutional neural network model, consisting of 138 million parameters, was utilised. This model achieved 92.7% top-5 test accuracy in ImageNet of 7.32% classification error, using only 3×3 convolution and max pool layers (16 layers in total), demonstrating that network depth leads to a higher classification accuracy. The built-in version of this network in PyTorch was adapted to allow for the input of the ultrasound images and the output of a binary vector classification. The first convolution layer was changed to allow 1 channel input images of size [1,58,52]. In addition, the last fully connected linear layer was changed to output a vector binary classification, and a final sigmoid layer was added.

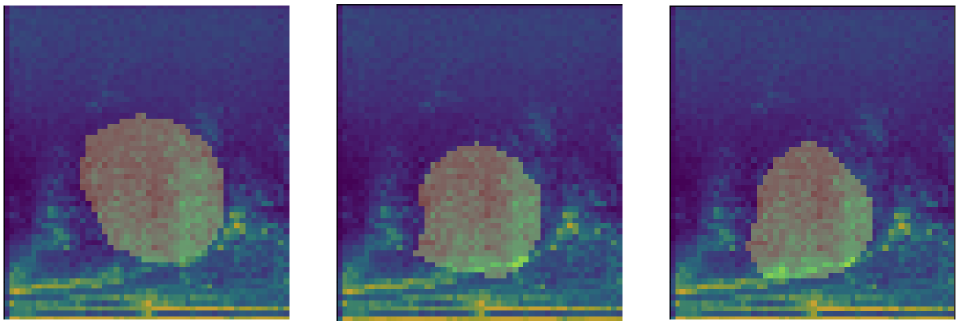
The training procedure of the VGG16 network follows closely that of the image segmentation process.The training data was once again augmented through the same geometrical transforms as with image segmentation, ensuring that the network learns to classify the presence of a prostate gland despite its orientation or position. The VGG16 model was trained for 500 epochs, where the binary cross entropy loss function was optimized at each epoch using the Adam optimizer. Once more, the average BCE loss and accuracy on training and validation data was outputted every epoch. The accuracy of the model was calculated as an average percentage of correctly predicted labels per epoch.

4 Experiments and Results

Firstly, based on the consensus segmentation labels as ground-truth, using an independent holdout test set, the two models for the two segmentation label sampling methods were compared. The test data used for comparison required a fixed set of cases, which was obtained by setting the seed of the random split of all data. In addition, this ensured that the two models were not trained on the test data. Furthermore, overlaid segmentation predictions for the two models were compared to the ground truth as shown in Fig 1. Secondly, the same process was repeated, after applying a screening using the trained VGG16 classification model. Once more, a holdout test set consisting of consensus segmentation labels and binary classifications was screened using the classifier. Frames where the prostate was not identified, were removed from the test data set. The screened test dataset was then used to test the accuracy of the two segmentation models and compared to the results from the previous experiment, as shown in Tab 1. Bland-Altman plots for the two segmentation methods with and without pre-screening is also shown in Fig 2.

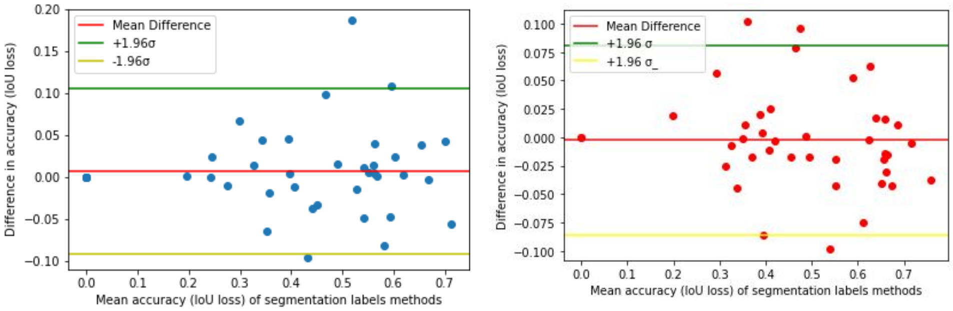
Segmentation method	Test loss	Test accuracy	Test loss	Test accuracy
Screening	No	No	Yes	Yes
Random	0.439	0.425	0.374	0.479
Consensus	0.449	0.414	0.374	0.481

Table 1: Tabular summary of IoU accuracy scores, after 500 epochs, for the two segmentation models with and without pre-screening classification applied.



(a) Ground truth (b) Segmentation 1 (c) Segmentation 2

Figure 1: Overlaid images of predicted segmentations and the ground truth on a frame after training for 500 epochs without pre-screening.



(a) Without pre-screening (b) With pre-screening

Figure 2: Bland-Altman plots comparing the accuracy of the two segmentation methods with and without pre-screening

5 Conclusion

To conclude, we have presented an UNet encoder-decoder architecture, made deeper by implementing ResNet, to segment prostate glands in transrectal ultrasound images. Using bootstrap aggregate and data augmentation, two models based on different label sampling methods were trained for 500 epochs and compared with each other. In addition the effectiveness of the model was further tested by applying pre-screening on test data using the VGG16 classification model. We notice that with pre-screening, the accuracy of the model increased, however fluctuations in accuracy between each epoch were noticed. It is also evident that there is no clear difference in accuracy between the two models. The segmentation of the prostate glands is promising when compared with the ground truth labels, but the model should be trained for a longer period of time to ensure better and more stable results.

References

[1] O. Ronneberger, P. Fischer and T. Brox, ”U-Net: Convolutional Networks for Biomedical Image Segmentation”, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp.234-241, 18 Nov 2015.

[2] He, K.: Deep Residual Learning for Image Recognition. CVPR (2015)

[3] Karen Simonyan, Andrew Zisserman, ”Very Deep Convolutional Networks for Large-Scale Image Recognition”, 4 Sep 2014

[4] Sebastian Ruder, ”An overview of gradient descent optimization algorithms”, 15 Sep 2016