Medical Ultrasound Classification and Segmentation

Thomas Ghorbanian[1]

University College London, Gower St, London WC1E 6BT, UK

1 Introduction

Deep learning models (UNet [1] , Gated-SCNN [2], FastFCN [3]) provide powerful alternatives to traditional image processing that allow for more flexibility in model design. Such models are adaptable for semantic classification as well as segmentation and can therefore lay the foundation for the development of end to end toolkits that can be used for computer vision tasks.

The following report proposes a supervised 2D image segmentation tool for transrectal B-mode ultrasound images of prostate gland capsules. A real-time model would have motivating applications in surgery, allowing surgeons to localise relevant anatomical structures and targeting regions of interest for various urologic procedures, such as ablation therapy and needle biopsies.

2 Methods

Ultrasound data from 200 cases of prostate gland imaging were prepared and labeled by three different urologists. Each case consists of multiple 58x52 single channel image frames and three labels, most of which are labelled as having a prostate gland capsule. Due to the inherent difficulty of identifying capsules, a supervised DenseNet [4] is used to learn detailed feature-maps to classify whether a given image frame contains prostate or not. A supervised UNet is then trained to learn the visual and positional feature-maps characterising prostate gland images to their corresponding labels.

2.1 DenseNet Design

Densely connected convolutional networks are designed to increase the depth of convolutional neural networks without loosing information characterising input images upon passing through many layers. By concatenating feature-maps learned from previous layers, the input of subsequent layers have greater variation in and improved efficiency.

A DenseNet-161 architecture from [4], accepting single channel 1x58x52 images and applying 6x6, 12x12, 36x36 and 24x24 dense block feature-maps, is used for binary classification. A final two dimensional softmax classifier is attached, such that outputs of $[1, 0]$ correspond to images containing prostate glands and

outputs of $[0, 1]$ correspond to those that do not. A growth rate of 32 is used to regulate how much new information each layer contributes to the global feature map.

2.2 UNet Design

An instance segmentation neural network combining the strengths of residual learning with UNet is designed to extract visual and spatial information from images [5]. The encoder is a convolutional network that consists of repeated application of convolutions, each followed by a rectified linear unit and a max pooling operation. During encoding, the spatial information is reduced while the feature information is increased. During decoding the features and spatial information are combined through a sequence of up-convolutions and concatenations.

A deeper alternative to double convolutions is implemented via residual blocks and connections [6]. Unlike DenseNets, ResNet adopt summation to connect preceding features (via residual skip connections) rather than concatenating them. Such a scheme eases training while retaining deep feature extraction. Residual blocks containing two 2-d convolutions followed by a batch normalisation and a rectified linear unit were used. The features from the previous block are added to the output of the next block. Four blocks are used in each residual network. The UNet expands the feature space from the initial single channel image to 32, 64, 128 and 512 channels respectively, before up transposing the other direction.

2.3 Loss and Accuracy

Loss The Dice distance between the predicted binary mask and the label, computed as 1 - $\mathcal{D}$ was minimized using the AdamW optimization algorithm. $\mathcal{D}$ can be computed as $\mathcal{D} = \frac{2|A \cap B|}{|A|+|B|}$, where $A \cap B$ represents the common elements between sets, and $|A|$ represents the number of elements in set A (and likewise for set B). The opted loss function for binary classification was the binary cross entropy.

Accuracy The Jaccard Index, $\mathcal{J}$, between the predicted binary mask and the label was opted to be used as the validation metric. The measurement emphasises the similarity between sample sets, and is formally defined as : $\mathcal{J} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$. Similarly to the Dice coefficient, its value can be interpreted as the percentage measurement of similarity between the two sample sets.

2.4 Regularisation

Data Augmentation Augmentation seeks to enforce invariance and robustness in the network, such as scale, shift and rotation invariance and deformation robustness. The strategy employed attempted to enforce shift and scale invariance (since the size and location of prostate glads varied significantly between

frames). To this end, image frames were randomly up-scaled in resolution and cropped back down to 58x52.

Bootstrapping Given the variability in predictions of retrained models, three bootstraps were trained for both the DenseNet classification and UNet segmentation to improve the stability. A consensus vote for the classifier (requiring 2> votes for a frame to be assigned a class) and a consensus vote at the pixel level for the segmentation (requiring 2> votes for a pixel to be considered as part of the segmentation) was used.

2.5 Data Sampling

Each frame had three binary mask segmentation labels from three independent observers. During training, each epoch sampled from all 200 cases without replacement, ensuring equal occurrence between cases. For each case, the dataloader randomly sampled one frame index without replacement, ensuring that all frames were equally likely to be sampled.

The labels were sampled using two methods. One method simply chose a random label index from the three available. The other method used a majority vote at the pixel level to form a consensus choice. In this scheme, the three labels were stacked and summed along the zeroth dimension. If the pixel values of the resulting matrix was above the halfway value of the number of labels, the pixel was replaced with a value of 1. Otherwise, a value of 0 was given.

3 Experiments

The ultrasound data was split into 120 training cases, 40 validation cases and 40 test cases. The training loop took steps on a minibatch of 16 training and validation images. Each bootstrap was trained for 100 epochs. The model was trained over three independent bootstraps, first training three classification models, and then feeding the images predicted to have a non-zero label into three independent segmentation models. Randomised sampling of image frames was used to train the independent bootstrap models, while pseudorandom (seeded) sampling was used to evaluate prediction averages between bootstraps (to keep index consistency when forming a consensus vote). A pure UNet model was trained to compare the two different segmentation label sampling methods.

Table 1. Table captions should be placed above the tables.

Model	Avg Train Accuracy	Avg Test Accuracy
UNet (random labels)	0.454	0.391
Unet (consensus labels)	0.450	0.414
DenseNet	0.810	0.700
Combined	0.847	0.670

4 Results

The performance of the models were assessed using the independent holdout test set, computing the mean Jaccard Index across all test cases. The results are shown in Table 1. A Bland-Altman plot was made comparing the two label sampling methods. An example segmentation is shown.

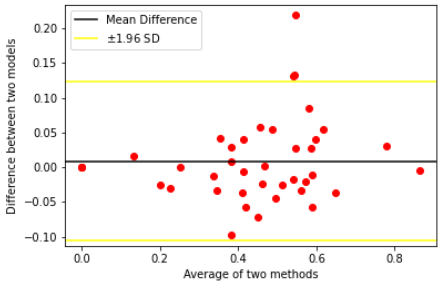
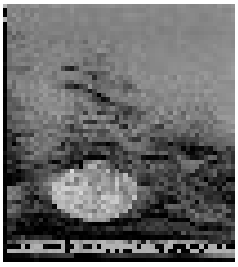
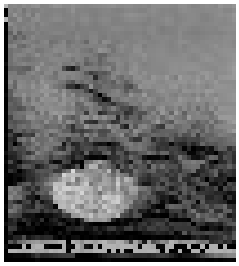


Fig. 1. Bland Altman plot comparing two sampling methods for pure UNet model. Sampling 1 is done via random label selection. Sampling 2 is done via consensus vote.



Ground truth image segmentation.



Predicted image segmentation. Jaccard index = 0.970

5 Conclusion

The best performance was achieved with the combined classification-segmentation model. Pre-screening dramatically improved the test accuracy over pure U-Net models. There was marginal improvement in segmentation when using consensus sampling over random sampling. A PyTorch implementation is provided, adaptable to be applied to different tasks.

References

1. Ronneberger, O.: U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI (2015)
2. Takikawa, T.: Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. ICCV (2019)
3. Wu, H.: FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. arXiv:1903.11816 (2019)
4. Huang, G.: Densely Connected Convolutional Networks. CVPR (2017)
5. Zhang, Z.: Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters (2017)
6. He, K.: Deep Residual Learning for Image Recognition. CVPR (2015)