Performance Evaluation of U-Net Convolutional Neural Network on Different Percentages of Training Data for Thyroid Ultrasound Image Segmentation

Prabal Poudel[1], Alfredo Illanes [1] and Michael Friebe[1]

Abstract— **Deep Learning has been a popular tool for various applications in the medical field. Segmentation of anatomical structures in ultrasound (US) images is an important medical image analysis tool for monitoring and diagnosis of several diseases. The use of US imaging is preferred compared to other imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI), nuclear imaging etc. as it is cheap with no harmful effects to the patients and medical practitioners. In this work, we segmented the thyroid gland in thyroid US images using a U-net architecture of convolutional neural network (CNN). We trained the U-Net CNN with different percentages of training data starting from 30% to 80%. We also evaluated the performance of the trained classifier on a test set in order to find out what percentage of training data are required to achieve a similar segmentation accuracy compared to other simple segmentation approaches like active contours without edges (ACWE), graph cut (GC) and pixel based classifiers (PBC). The maximum Dice Coefficient of 0.896 was obtained when the the U-Net was trained with 80% of training data on a thyroid dataset with 1600 2D US images.**

I. INTRODUCTION

The thyroid is one of the largest endocrine glands in the human body which is involved in several body mechanisms like controlling energy sources usage, synthesis of proteins and controlling body's sensitivity to other hormones. It is important to keep track of the thyroid shape and size over time as it is susceptible to many diseases like Graves, sub-acute thyroiditis, thyroid cancer, etc. Segmentation and volume computation are thus essential to monitor the thyroid shape and size over time.

Several thyroid US image segmentation approaches based on classic image processing techniques such as ACWE, GC and PBC [1] have been proposed. Augustin *et al.* [2] used a fuzzy c-means algorithm, histogram clustering, QUAD tree, region growing and random walker methods to segment thyroid US images. A novel boundary detection method and local binary patterns for texture analysis in thyroid US images were proposed in [3]. Similarly, several supervised learning based thyroid segmentation techniques have been proposed. A polynomial support vector machine (SVM) to segment the thyroid gland in US images was used by Selvathi *et al.* [4]. Garg *et al.* [5] used a feedforward neural network to segment thyroid glands in US images. A radial basis function (RBF) neural network to segment the blocks of the thyroid gland was used by Chang *et al.* [6].

These approaches are mainly based on computing different statistical features such as mean, variance, entropy, histograms, root mean squared differences, etc. in thyroid US images and later using these features to differentiate the thyroid from the non-thyroid regions. Despite computation of these many features, it is a challenging task to segment the thyroid in US images due to the presence of speckle noise, several artifacts as well as a low signal to noise ratio and resolution. Even the filtering methods cannot get rid of these noises completely. Hence, a method that can extract highly detailed and extensive features at pixel level and is not affected by the presence of speckle noise can only achieve a good thyroid segmentation accuracy. Similarly, all the machine learning (ML) approaches in the literature too require a large amount of training data to achieve a good segmentation result. On top of that, it is a difficult task to choose the right amount of training and testing data to achieve this accuracy.

Hence, in this work, we train our U-Net with different percentages of training data and compare the segmentation accuracy of U-Net with different non-machine learning based methods such as ACWE, GC and PBC. The main purpose of this paper is not to propose a new segmentation algorithm but to find out what amount of data are required by U-Net to achieve the same level of segmentation accuracy compared to these non ML based methods which require handcrafted features, image pre-processing and on top of that, are not automated.

II. METHODS AND PROCEDURE

A. Thyroid Ultrasound Dataset

We have used two datasets in this work to evaluate the proposed approach. The first dataset (Dataset 1) consists of six subjects with each subject containing between 53 and 189 2D thyroid US images making it a total of 675 images with an image size 760 x 500 pixel. This dataset has already been published and is available in [7]. Similarly, the second dataset (Dataset 2) was published in [8] and can be downloaded from `http://opencas.webarchiv.kit.edu/?q=node/29`. This dataset contains images from 16 different subjects with each subject containing 100 2D thyroid images and were acquired by a different physician than in the first dataset. The second dataset contains a total of 1600 2D US images with an image

*This work has been funded by Federal Ministry of Education and Research in the context of the INKA Project. [Grant Number 03IPT7100X]

[1]Prabal Poudel, Alfredo Illanes and Michael Friebe are with Faculty of Medical Engineering, Otto-von-Guericke University, Magdeburg, Germany `prabal.poudel@ovgu.de`

Manuscript 895 submitted to 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Received February 4, 2019.

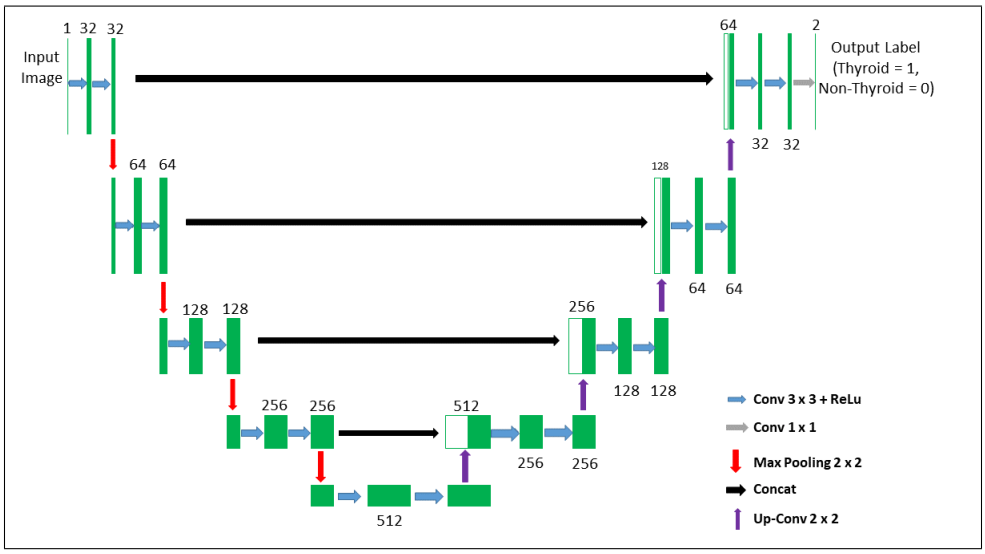

Fig. 1. Architecture of the U-Net CNN

size of 760 x 1020 pixel. Along with the US images, we also acquired manually segmented ground truth images from the same clinical experts. All the images were acquired using a General Electric (GE) Logiq E9 US machine equipped with an electromagnetic tracking system.

B. Segmentation using U-Net CNN

In this work, we have employed the U-Net architecture proposed by Ronneberger *et al.* [9]. Fig. 1 shows the network architecture which consists of the down-sampling and up-sampling parts that analyses the images by contracting in each successive layers and then expanding in order to produce a full-resolution segmentation respectively. The input to the CNN was $X = (I_n, G_n)$, where I_n represents a thyroid US image and G_n represents its ground truth.

The down-sampling path consists of 3 x 3 convolution followed by a rectifier linear unit (ReLu) in each layer and then a 2 x 2 max pooling with no stride. The output feature space is doubled in each layer in the down-sampling path and the up-sampling path remaps the lower resolution feature maps to a higher resolution space of the input images. This is done by up-sampling the feature maps followed by a 2 x 2 convolution (up-convolution) which halves the number of feature channels in each up-sampling step, a concatenation with the corresponding feature map from the down-sampling path and two 3 x 3 convolutions each followed by a ReLu activation. The final activation function predicts the output label for all the pixels in the image(i.e. thyroid or non-thyroid as the output of the network). 'Binary Crossentropy' was set as the loss function and 'adam' optimizer was used for the minimization of the loss function and the CNN was trained for 20 epochs.

Data augmentation was carried out by flipping (because the thyroid gland has two lobes) and scaling the images (because different subjects have different thyroid sizes) and varying the lighting condition (to simulate the gain parameter in the US machine during the image acquisition). The learning rate of the network was set to be 1 x 10^{-5}. In order to avoid the problems of over-fitting, we added a dropout of 0.25 in both the down-sampling and up-sampling layers after the first convolution in each layer. Additionally, all the images were normalized using the Z-score method (i.e. subtracting the mean and then dividing by the standard deviation of the pixel intensities of each patch).

C. Post-Processing

The output of the CNN was a label (i.e. thyroid = 1 and non-thyroid = 0) for each pixel in an US image. This produces a segmented image which is a binary image. The segmented images were undergone a post-processing stage. In this stage, we performed a largest connected component analysis. The total number of thyroid pixels were obtained by counting the pixels that were classified as thyroid (i.e. output label = 1). Some segmented images contained multiple blocks or patches (i.e. over segmentation) of thyroid pixels. Hence, only the block of pixels containing more than 60% of the total thyroid pixels were considered as thyroid region while the blocks of pixels containing less than 60% of thyroid pixels were disregarded.

III. EXPERIMENTAL RESULTS

For the evaluation and quantitative and qualitative analysis of the approach for thyroid segmentation using U-Net CNN, we performed a series of experiments with different percentage of training and testing data. These experiments included training the U-Net with 30%, 40%, 50%, 60%, 70% and 80% of training data from the entire database. All the experiments were carried out using a Lenovo T430 ThinkPad Notebook with Intel Core i5-3320M CPU, NVIDIA NVS 5400 graphics card, 2.60 GHz processor and 8.00 GB RAM.

Two sets of tests were performed, one with the Dataset 1 and another with Dataset 2. The final results of segmentation in a thyroid US image using U-Net with different training percentages using both datasets are presented in Fig. 2. In

Manuscript 895 submitted to 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Received February 4, 2019.

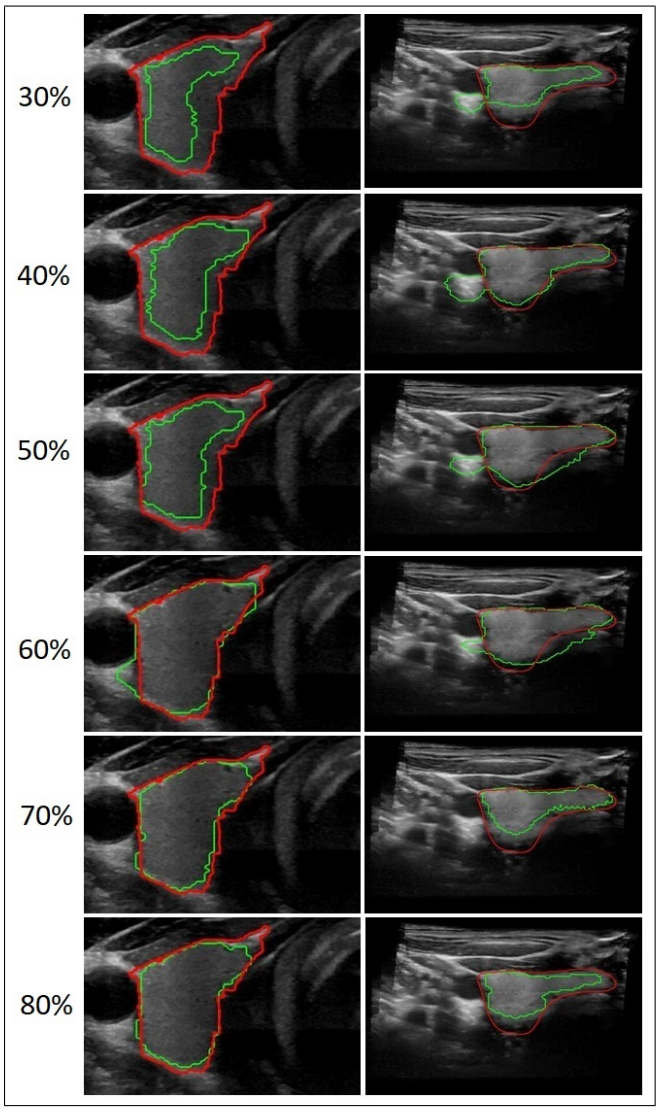

Fig. 2. Segmented thyroid (green) compared with ground truth (red) using different percentages of training data starting from 30% in the first row to 80% in the last row with an increment of 10 % in each row. The columns represent the images from Dataset 1 and 2 respectively.

the figure, the solid red line represents the ground truth and the solid green line represents the boundary of segmented thyroid. Similarly, the quantitative results are presented in Table I and II respectively for Dataset 1 and 2. The tables present the DC for different percentages of training data. In the tables, TT means Training Time (in Sec), TA means Test Accuracy (in DC) and (%) means percentage of training images.

Some more results of segmentation using U-Net from different subjects and from different locations with respect to thyroid volume is shown in Fig. 3. In the figure, the first row represents the results using Dataset 1 and similarly the second row using Dataset 2. The results show that with less amount of training data, the segmentation results are not highly accurate while, increasing the amount of training data improves the segmentation results.

In order to compare the segmentation results of U-Net

TABLE I

SEGMENTATION RESULTS USING DATASET 1

(%)	Training	Testing	TT	TA
30	211	492	692	0.685
40	281	422	923	0.707
50	352	351	1134	0.734
60	422	281	1372	0.804
70	492	211	1609	0.837
80	562	141	1835	0.872

TABLE II

SEGMENTATION RESULTS USING DATASET 2

(%)	Training	Testing	TT	TA
30	480	1120	1575	0.706
40	640	960	2065	0.748
50	800	800	2567	0.807
60	960	640	3312	0.838
70	1120	480	3662	0.861
80	1280	320	4181	0.896

with other approaches, we implemented three classic non-machine learning based approaches, ACWE, GC and PBC [1]. Out of the 10 subjects tested in [1], we only compare the results with six subjects (which are from Dataset 1) in this work. Similarly, we take only ACWE, GC and PBC out of the five approaches of [1] in this work for comparison since these were evaluated on 2D image datasets while the rest two approaches were evaluated on the 3D US image datasets. The comparison is done based on DC and HD and is presented in Table III.

ACWE starts with the user drawing an initial rectangular or square contour around the thyroid region and the initial contour evolves over time to produce a final segmented thyroid. Similarly, in GC, the user marks the thyroid and non-thyroid regions using different coloured scribbles. The user marked regions are used to construct different Gaussian Mixture Models (GMMs) which ultimately assign each pixel to either thyroid or non-thyroid region. PBC allows the user to click inside the thyroid and non-thyroid region from where different features are computed. These features are used to train several decision trees which classify each pixel into thyroid and non-thyroid. Dice Coefficient (DC) and Hausdorff Distance (HD) were used as the performance measures to compare the segmentation results between U-Net and the approaches from [1].

Similarly, we compare our results with [10] which uses all the 16 subjects from Dataset 2 to segment the thyroid using Iterative Random Walks and Random Forest (IRWRF). In their work, they have compared IRWRF with Echogenicity-based Quantization (EBQ), Joint Classification-Regression (JCR), Radial Basis Function (RBF) Neural Network and Feedforward Neural Network (FNN). The comparison is made based on the DC, Sensitivity (SE) and specificity (SP) and presented in Table IV.

The results from Table I and III show that U-Net requires less than 30% of training data to achieve the same accuracy

Manuscript 895 submitted to 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Received February 4, 2019.

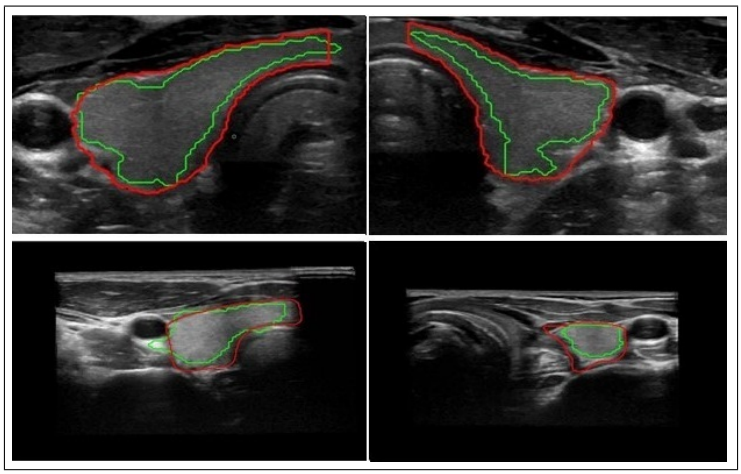

Fig. 3. Segmented thyroid (left and right lobes) from 2 different subjects in Dataset 1 (first row) and Dataset 2 (second row)

TABLE III
COMPARISON OF U-NET WITH [1] IN TERMS OF DC AND HD USING DATASET 1

	ACWE	GC	PBC	**U-Net**
DC	0.805	0.745	0.667	**0.872**
HD	8.1	8.3	9.5	**7.1**

TABLE IV
COMPARISON OF U-NET WITH [10] IN TERMS OF DC, SP AND SE USING DATASET 2

	IRWRF	EBQ	JCR	RBF	FNN	**U-Net**
DC	0.854	0.839	0.479	0.512	0.400	**0.867**
SP	92.3%	88.9%	92.6%	56.0%	86.4%	**83.4%**
SE	98.9%	95.5%	56.4%	87.4%	47.3%	**89.2%**

as PBC, around 50% of data to achieve the same accuracy as GC and 60% of training data to achieve the same segmentation accuracy compared to ACWE. This proves the efficiency of U-Net compared to classic non-machine learning approaches and on top of that U-Net is fully automatic and when the training amount is increased, provides better segmentation results.

IV. CONCLUSIONS

In this work, we implemented U-Net CNN approach to segment 2D thyroid US images. Along with the segmentation, we performed a series of experiments using different percentages of training data to find out what amount of training images are required for CNN to achieve the same segmentation accuracy as compared to some of the non-machine learning based approaches (i.e. ACWE, GC and PBC). We also showed that, on increasing the amount of training data, the segmentation accuracy increases which proves that the training stage in CNN is very important and a right amount of training data should be chosen. Furthermore, we showed that CNN is robust, does not require any pre-processing step and is fully automatic unlike the compared approaches. Thus, this method can be used compared to the non-automated and time consuming non-machine learning based approaches for image segmentation and many more medical image processing tasks.

ACKNOWLEDGMENT

We would like to thank General Electrics, USA for providing us with the LogiqE9 Ultrasound Machine to generate Thyroid Ultrasound data. Special thanks to our clinical partner at the University of Magdeburg (Prof. C. Arens) for helping us obtain the Thyroid Ultrasound Datasets.

REFERENCES

[1] P. Poudel, A. Illanes and M. Friebe, Evaluation of Commonly Used Algorithms for Thyroid Ultrasound Images Segmentation and Improvement Using Machine Learning Approaches. Journal of Healthcare Engineering, 2018.

[2] A. S. Augustin, S. S. Babu and K. T. Nadu, Thyroid segmentation on us medical images: An overview, 2012.

[3] E. G. Keramidas, D. K. Iakovidis, D. Maroulis and S. Karkanis, Efficient and effective ultrasound image analysis scheme for thyroid nodule detection, Springer, pages 1052-1067, 2007.

[4] D. Selvathi and V. S. Sharnitha, Thyroid classification and segmentation in ultrasound images using machine learning algorithms, in Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on. IEEE, pp. 836-841, 2011.

[5] H. Garg and A. Jindal, Segmentation of thyroid gland in ultrasound image using neural network, in Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, pp. 1-5, 2013.

[6] C. Y. Chang, Y. F. Lei, C. H. Tseng and S. R. Shih, Thyroid segmentation and volume estimation in ultrasound images, IEEE transactions on biomedical engineering, vol. 57, no. 6, pp. 1348-1357, 2010.

[7] P. Poudel, E. Ataide, A. Illanes, M. Friebe, Linear Discriminant Analysis and K-Means Clustering for Classification of Thyroid Texture in Ultrasound Images. Proc IEEE Eng Med Biol Soc, 2018.

[8] T. Wunderling, B. Golla, P. Poudel, C. Arens, M. Friebe, C. Hansen, Comparison of thyroid segmentation techniques for 3D ultrasound. Proc. SPIE Med. Imaging, 2017.

[9] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234-241, 2015.

[10] D. China, A. Illanes, P. Poudel, M. Friebe, P. Mitra, D. Sheet, Anatomical Structure Segmentation in Ultrasound Volumes using Cross Frame Belief Propagating Iterative Random Walks. IEEE journal of biomedical and health informatics, 2018.