Magnetic Resonance Sequences: Experimental Assessment of Achievements and Limitations

Pedro Furtado
DEI-CISUC, Universidade de Coimbra, Coimbra, Portugal
Email: pnf@dei.uc.pt

***Abstract*—Deep Learning can be applied to learn segmentations of abdominal organs in MRI sequences, a challenging task due to changing morphologies of organs along different slices. Evaluation of outcome is important to decide on applicability and to command further improvements. Software tools include evaluation metrics. Some metrics indicate quasi-perfection, with potential erroneous conclusions, visual inspection and some per organ metrics say otherwise. Our aim is the correct interpretation of commonly available metrics on organs segmentation. The method to do that is to build two architectures (DeepLab, FCN), run segmentation experiments, interpret results. Examples of results as aggregates (mean accuracy 98% weighted IoU 97%) are overly optimistic. Further analysis shows much lower scores (mean IoU 68% IoU of individual organs 78, 66, 59, 41%). We conclude that correct interpretation of the metrics, importance of further architectural or post-processing improvements on false positives.**

***Index Terms*—segmentation, deep learning, assessment**

I. Introduction

Magnetic Resonance Imaging (MRI) is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body. MRI scanners use strong magnetic fields, magnetic field gradients, and radio waves to generate images of the organs in the body [1]. MRI acquires sequences of body part slices that are transformed into images useful for visualization and analysis by medical doctors. It is possible then to study the images and to reconstruct whole structures using the sequence of images. This allows medical doctors to diagnose medical conditions based on their analysis.

Automatic segmentation of structures in MRI images is an optional add-on that can help highlighting specific structures or be used in subsequent analysis steps. In recent years new deep-learning approaches to segmentation made their way into medical imaging with astonishing success. By training with a large set of images and corresponding segmentations, Deep Convolution Neural Networks (DCNN) with decoding capability gain the ability to classify each image pixel of any new image as being part of a specific structure. This new technology was made possible by a set of innovations that includes convolution layers and backpropagation learning. After much work developing and improving DCNN architectures for the task of segmentation, the focus is now also on applying those technologies in specific imaging tasks and evaluating their performance on those tasks. Chaos [1] and [2] is a medical imaging challenge related to segmentation of abdominal organs and submitted to the International Symposium on Medical Imaging (ISBI) 2019. The purpose of the challenge is to allow understanding of prerequisites of complicated medical procedures. As part of the challenge, competitors are asked to develop and test techniques for the segmentation of four abdominal organs (i.e. liver, spleen, right and left kidneys) from Magnetic Resonance Imaging (MRI) data sets acquired with two different sequences (T1-DUAL and T2-SPIR). Results of the challenge are available in [3], on a scale 0-100 the leader board varied between 25.6 and 66.4. The leaderboard values use a mix of evaluation metrics described in [4], with the results of four metrics converted to grades at 0-100 scale with help of pre-defined thresholds.

Today, many deep learning architectures have already been developed and tested, and there are software toolkits available for application in this kind of segmentation problems. Those toolkits also include segmentation performance evaluation metrics that should be used to verify whether the quality of the learned network is sufficiently good to be used in practice. However, it is very easy, based on some of those metrics and the way they are captured, to take for granted that the approaches were able to segment with almost 100% quality, while a simple visual inspection of some images already reveals deficiencies in correctly outlining some of the organs. This prompted our investigation into experimenting with state-of-the-art approaches on the task of segmentation of abdominal organs and studying the correct use and interpretation of metrics. This is a very important step for the correct evaluation of approaches and to determine what is still needed in future work.

A. Related Work

The Convolution Neural Network (CNN) is a specialized type of neural network with a large number of convolution layers that extracts and sequentially abstracts features from images, with the first layers operating in more restricted local fields of view of those images and the subsequent ones abstracting those extracted features.

Manuscript received June 14, 2020; revised November 23, 2020.

doi: 10.12720/jait.12.1.66-70

Th purpose of the CNN is to classify the images, which is done by a fully connected artificial neural network put at the end of the CNN. Backpropagation learning allows the CNN to adjust all its filter and neuron weights to improve classification based on training images and corresponding classes. The segmentation network is also a CNN but instead of classifying images it classifies each individual pixel of the images as belonging to one class, in effect doing semantic segmentation of the image. Fully Convolutional Network (FCN) were proposed in [5] as an effective approach for segmenting images using this logic. The segmentation network is built by using a CNN as encoder network, removing the fully connected network at the end and replacing it by a sequence of deconvolution or upsampling layers until the original image size is restored. FCN can be fitted with any CNN, we apply VGG16 [6]. FCN scores more than 62% on the Intersect-over-Union (IoU) metric on the 2012 PASCAL VOC segmentation challenge. U-Net [7] followed, as a DCNN well-fitted for segmentation of medical images. Finally, DeepLab [8] introduced Atrous Spatial Pyramid Pooling (ASPP) to capture objects at various scales, and also probabilistic graphical models that improve object boundaries. It scored 80% on IoU PASCAL VOC-2012. Our implementation of FCN uses Resnet-18 as the CNN.

In what concerns segmentation of abdominal organs, an example of a traditional approach using texture segmentation is the proposal in [9]. More recently, researchers have been proposing segmentation using deep learning. For instance, [10] proposes improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function, and [11] shows that fully convolutional neural networks improve abdominal organ segmentation when compared with multi-atlas method. The FCNN resulted in a Dice Similarity Coefficient (DSC) of 0.930 in spleens, 0.730 in left kidneys, 0.780 in right kidneys, 0.913 in livers, and 0.556 in stomachs. The performance measures for livers, spleens, right kidneys, and stomachs were significantly better than multi-atlas alternative. In another work, [12] proposes a method for reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections.

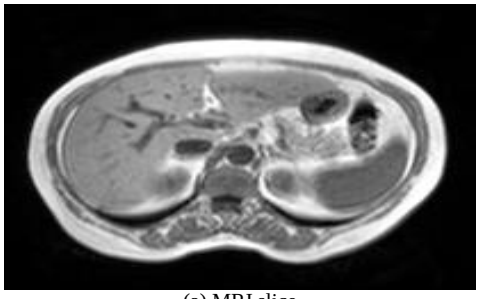
Metrics are also very important in the evaluation of the approaches, [13] and [14] already point the way to correctly evaluate segmentation of biomedical images. Among the conclusions, and considering only the simplest metrics, we can say that some metrics that include precision and recall or IoU can give more accurate assessment than accuracy, but further considerations are needed to complete the discussion on metrics, which we do in the next section.

II. Methods of Investigation

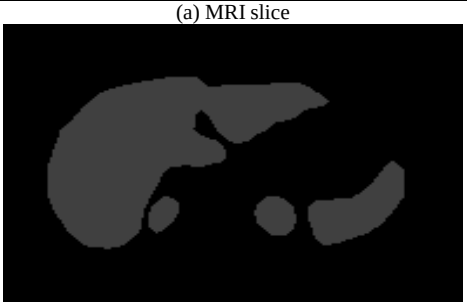
In order to investigate the problem raised in the introduction, we built state-of-the-art deep-learning segmentation architectures, trained them to segment abdominal organs in the CHAOS dataset, evaluated their worth according to typical performance metrics and using also visual inspection, then we studied the interpretation of those metrics to show how to correctly characterize the quality of the approaches. That includes explaining why metrics show values that are somehow apparently contradictory with visual inspection and with other metrics, and settling to an interpretation of the available metrics that allows us to reveal the deficiencies in the segmentation outcomes, as much as the qualities. In the following we discuss the materials used.

A. Dataset

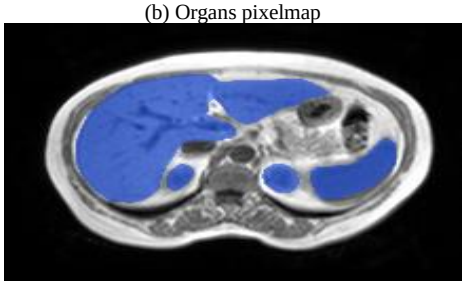
The experiment of semantic segmentation was to take any MRI slice from a dataset, such as Fig. 1(a) and obtain a correct pixel map as in Fig. 1(b) automatically. Fig. 1(c) shows the pixelmap overlapped on top of the image.



(a) MRI slice



(b) Organs pixelmap



(c) Pixel map superimposed

Figure 1. Example of MRI slice, pixel map and superimposed image (Chaos [2]).

A 1.5T Philips equipment was used to obtain MRI sequences of the abdominal region including the kidneys, spleen and liver organs, resulting in 12-bit DICOM images, 256×256 resolution. The ISDs varies between 5.5-9 mm (average 7.84 mm), x-y spacing is between 1.36 - 1.89 mm (average 1.61 mm). Acquisition used T1-DUAL modality, a fat suppression sequence using difference of T1 times between fat and water protons. 1594 slices were acquired and used in the experiments (each sequence had approx. 532 slices), the patient sequences were analyzed using 5-fold cross-validation (5 runs, 80/20 train-test divisions) and average values from the 5 runs are reported.

B. Techniques Built and Compared

Our experiments involved training and testing two state-of-the-art DCNNs for image segmentation, DeepLabV3 [8] and FCN [5] on the MRI sequences. The networks and experimental setup were implemented in Matlab2019a, and the networks were modified to balance class weights. The following initial training options were used.

Type of learn rate schedule = 'piecewise' (rate decreases after period)

Learn rate drop period=10 (time to decrease), factor=0.8 (how much), Momentum=0.9, Initial learn rate= 0.001;

Maximum number of epochs= 500, batch size= 8, Shuffle every epoch,

Plot the training progress, stop only at the end of epochs or by user command.

C. Metrics

Segmentation performance metrics are essential to evaluate the approaches correctly, but the correct use and interpretation of those metrics commonly provided by toolkits is especially important. For instance, it is possible to have 98% overall accuracy, but only 20% IoU segmenting a kidney. Without a correct interpretation of the metrics, erroneous conclusions can ensue. Having a single numeric quantity (e.g. accuracy between 0 to 100%) to qualify accuracy is very attractive because it lends itself to ease of comparison, but it usually masks relevant segmentation deficiencies that can be revealed in a more detailed analysis. We propose that two main dimensions be considered when analyzing an approach based on performance metrics. One dimension is the nature of the metric itself (e.g. Accuracy or Intersection over Union), the other also very important dimension is the scope. The scope defines whether the metric is captured globally, for all pixels, per class (e.g. evaluated for each organ), per image or both. We will show that a correct evaluation requires a set of metrics in terms of both nature and scope, and that it is fundamental to understand why a metric exhibits a certain value, the interpretation of the metric.

III. Analysis of Experimental Results

The approaches were trained according to the experimental setup described in the previous section on randomly selected 1275 slices, while the remaining slices were used for testing. In this section we present the experimental results, organized into different sections showing different metrics and the visual inspection of some resulting segmentations, then we analyze the interpretation of those results that reveals the true nature (quality and deficiencies) of the approaches.

A. Global Performance and Confusion Matrices

Table I and Table II show global accuracy and confusion matrices. These results will be analyzed in detail in the next section. Table I shows the global accuracy metrics (scope = global, nature=accuracy, IoU and BFScore). Note the difference between the values from the first three columns (Global Accuracy, Weighted IoU and Mean Accuracy) and the other two columns (Mean IoU and Mean BFScore). Table II and Table III show the confusion matrices for DeepLabV3 and FCN respectively as reported by Matlab2019 toolbox function 'evaluateSemanticSegmentation', which normalizes by the number of pixels known to belong to each class. Note that all classes (organs plus background) have very good true positive rates in the confusion matrices.

TABLE I. Global Accuracy

	Global Accuracy	Weighted IoU	Mean Accuracy	Mean IoU	Mean BFScore
DEEPLABV3	**0.98**	**0.97**	**0.98**	**0.69**	**0.74**
FCN	0.97	0.96	0.91	0.54	0.61

TABLE II. Confusion Matrix DeepLabV3

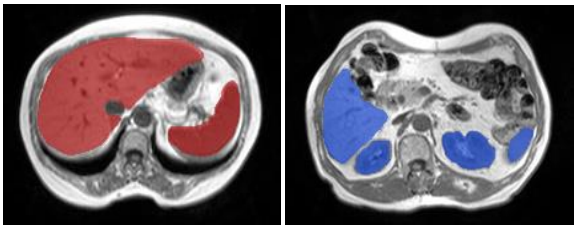
DEEPLAB	BackGrd	liver	spleen	rkidney	lkidney
BackGrd	**0.98**	0.01	0.00	0.00	0.01
liver	0.03	**0.97**	0.00	0.00	0.00
spleen	0.01	0.01	**0.98**	0.00	0.00
rkidney	0.04	0.00	0.00	**0.95**	0.01
lkidney	0.01	0.00	0.00	0.00	**0.99**

TABLE III. Confusion Matrix FCN

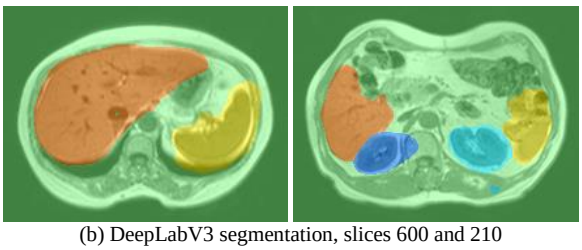
FCN	BackGrd	liver	spleen	rkidney	lkidney
BackGround	**0.97**	0.01	0.00	0.01	0.01
liver	0.07	**0.91**	0.01	0.01	0.01
spleen	0.01	0.03	**0.91**	0.05	0.00
rkidney	0.05	0.01	0.01	**0.89**	0.02
lkidney	0.04	0.07	0.00	0.01	**0.87**

B. Inspection of Some Slices

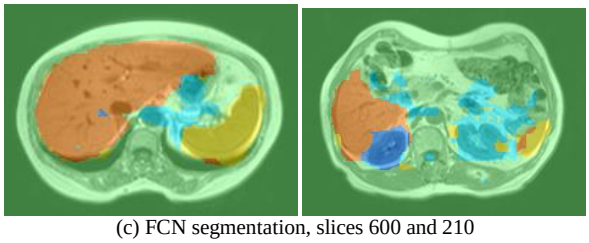
Fig. 2 shows two slices of the MRI (left and right columns respectively), with (a) showing the groundtruth segmentations, (b) showing DeepLabV3 segmentations and (c) showing FCN segmentations. Note a lot of imperfections. Fig. 3 shows the same for slice 71, and once again it is possible to see a lot of imperfections.



(a) Organs in slices 600 and 210: groundtruth

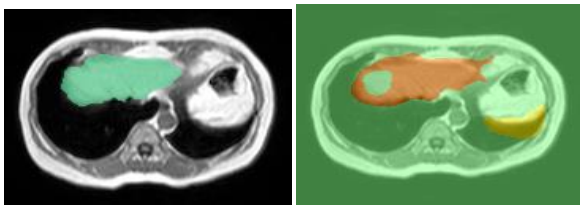


(b) DeepLabV3 segmentation, slices 600 and 210

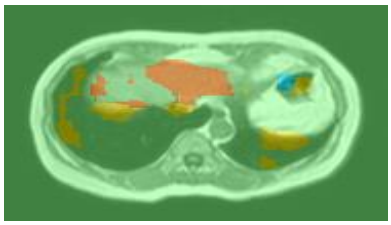


(c) FCN segmentation, slices 600 and 210

Figure 2. Inspection of slices 210 and 600.



(a) Groundtruth (b) DeepLabV3



(c) FCN

Figure 3. Inspection of slice 71.

C. Per-class Performance

Table IV and Table V (DeepLabV3 and FCN) show per-class accuracy (scope = per-class, nature=accuracy, IoU and BFScore). Note the difference between the first column, accuracy, and the last two columns, and how class "background" has very results for all metrics shown in the tables.

TABLE IV. PER-CLASS PERFORMANCE DEEPLABV3

DEEPLABv3	Accuracy (%)	IoU (%)	BFScore (%)
BackGrd	97	97	90
liver	96	78	60
spleen	97	59	56
rkidney	94	66	56
lkidney	98	41	36

TABLE V. PER-CLASS PERFORMANCE FCN

FCN	Accuracy (%)	IoU (%)	BFScore (%)
BackGround	0.96	0.96	0.81
liver	0.90	0.68	0.46
spleen	0.90	0.49	0.39
rkidney	0.88	0.18	0.18
lkidney	0.86	0.35	0.36

D. Analysis and Interpretation of the Metrics

Looking at the results shown in the previous subsections, it becomes clear that there is a need for a careful interpretation of the metrics to explain the wide amplitude of values shown and consequently to evaluate the approaches correctly. For instance, while the accuracy of DeepLabV3 is 98% (Table I), its IoU on the left kidney (Table V) is only 35%, and on the right kidney it is only 18%. Note also that the same Table V shows that segmenting the right kidney has an accuracy of 88%.

Starting the analysis by inspection of Fig. 2 and Fig. 3, it shows that DeepLabV3 was "almost perfect" segmenting the liver in slice 600, but it was not as perfect on the spleen, and in the remaining slices there are more evident deficiencies of DeepLabV3 segmenting organs. FCN had an even much larger amount of deficiencies. These serve to illustrate that the accuracies of 98% and 97% in Table I do not reflect the problems seen in the inspections.

Analyzing Table I, global accuracy, mean accuracy and weighted IoU are all very high (>96%). But the last two metrics in Table I (e.g. 54% meanIoU for FCN, 69% for DeepLabV3) are much lower. The reason for the discrepancy is that more than 95% of all pixels in all slices are background, therefore metrics over all pixels reflect almost entirely the capacity to segment the background, and the overall majority of background pixels are well segmented. A more detailed analysis of the formulas based on True Positives (TP), False Negatives (FN) and False Positives (FP) reveals why even mean accuracy (Table I), per-class accuracy (in Table IV and Table V) or the confusion matrices (Table II and Table III) mask defects significantly. The formula for accuracy or each class is acc(c) = TPc/(TPc+FNc), where c stands for the class. This formula does not include False Positives (FPc) that would reveal "spillovers" from organs segments to the background. Here we are defining spillovers as incorrectly taking background pixels as an organ. Those "spillovers" are revealed only as FNc in acc(background), but there the TPc area is so big that acc(background) is also still around 97%.

Contrary to accuracy, per-class IoU and BFScore in Table IV and Table V help clarify that organs are not very well segmented. For instance, the IoU measurements of the best-performing technique (DeepLabV3) are (97%, 78%, 59%, 66% and 41%) for classes (background, liver, spleen, rkidney, lkidney) respectively, revealing that only the background is very well segmented. Intersection over the Union (IoUc) reveals the problems much better than per class accuracy because of the term FPc in IoU(c)= TPc / (TPc + FNc + FPc), which finds the spillovers. Note also that BFScore, which measures the degree of matching between actual and predicted boundary (being within a predefined distance) also reveals difficulties segmenting organs adequately, and it is strongly positively correlated with IoU. Finally, the mean IoU and mean BFScore from Table I already reveal a bit more than accuracy, but since they are mean values and class background pushes them up, it is better to consider metrics over individual organs.

While the confusion matrices of Table II and Table III are normalized by the number of pixels known to belong to each class, in Table VI we show the non-normalized version for DeepLabV3, which reveals a large number of false positives for each organ.

TABLE VI. CONFUSION MATRIX, DEEPLABV3

	BackGround	liver	spleen	rkidney	lkidney
BackGround	6,332,500	36,857	10,898	8,842	44,277
liver	4,704	158,170	180	0	134
spleen	175	233	17,251	0	0
rkidney	733	0	1	19,907	230
lkidney	230	0	0	16	32,093

IV. CONCLUSION AND FUTURE WORK

Ready-to-use toolkits for development and testing of deep learning-based segmentation make available a usual set of metrics, in this paper we have studied the correct interpretation of those metrics in the specific context of segmentation of abdominal organs from MRI sequences. We have shown that some metrics do not reveal deficiencies in segmentation of the organs sufficiently, it is necessary to analyze per-organ metrics, to interpret all metrics and to include IoU metrics as well. This also shows there is important space for future improvements in deep segmentation procedures. Our future work on the subject includes post-processing steps based on properties of the regions to improve the outcome.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

This work has been entirely conceived, supervised, done and written by the sole author, Pedro Furtado. Pedro Furtado supervised all the work, conducted the research, analyzed the data, wrote the paper and approved the final version.

ACKNOWLEDGMENT

We would like to thank the CHAOS challenge organizers for sharing the dataset [2] and [3], since it made it possible for us to do this work, and for other researchers to advance the field as well.

REFERENCES

[1] Challenge. (December 2018). CHAOS - Combined (CT-MR) healthy abdominal organ segmentation. *Computer Vision News Magazine*. RSIP Vision. [Online]. Available: Available: https://www.rsipvision.com/ComputerVisionNews-2018December/16/

[2] CHAOS data DOI number now. [Online]. Available: https://doi.org/10.5281/zenodo.3362845

[3] Chaos results. [Online]. Available: https://chaos.grand-challenge.org/ Results_CHAOS/

[4] Chaos evaluation metrics. [Online]. Available: https://chaos.grand-challenge.org/Evaluation/

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241.

[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2017.

[9] J. Wu, S. Poehlman, M. Noseworthy, and M. Kamath. "Texture feature based automated seeded region growing in abdominal MRI segmentation," presented in International Conference on BioMedical Engineering and Informatics, Sanya, China, 27-30 May 2008.

[10] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," arXiv preprint arXiv:1707.04912, 2017.

[11] M. Bobo, S. Bao, Y. Huo, Y. Yao, J. Virostko, A. Plassard, and B. Landman, "Fully convolutional neural networks improve abdominal organ segmentation," in *Proc. SPIE Int. Soc. Opt. Eng.*, Mar. 2018.

[12] G. Chlebus, H. Meine, S. Thoduka, N. Abolmaali, B. Ginneken, H. Hahn, and A. Schenk, "Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections," *PLoS ONE*, vol. 14, no. 5, 2019.

[13] E. Tiu. (2019). Metrics to evaluate your semantic segmentation model. [Online]. Available: https://towardsdatascience.com /metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2

[14] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?" in *Proc. the British Machine Vision Conference*, 2013, pp. 32.1-32.11.

Pedro Furtado Pedro Furtado is Professor at University of Coimbra UC, Portugal, teaching Computer and Biomedical Engineering, with more than 25 years' experience. Pedro currently focuses in biomedical image and data analysis, with more than 200 papers published in international conferences and journals. Besides a PhD in Computer Engineering from U. Coimbra (UC) (2000), Pedro Furtado also holds an MBA from Universidade Catolica Portuguesa (UCP) (2004). Pedro is a member of SPIE.