Organ Segmentation in CT Images With Weak Annotations: A Preliminary Study
Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Fahed Abdallah

To cite this version:
Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Fahed Abdallah. Organ Segmentation in CT Images With Weak Annotations: A Preliminary Study. 27-ème Colloque GRETSI sur le Traitement du Signal et des Images, Aug 2019, Lille, France. hal-02183031

HAL Id: hal-02183031
https://hal-normandie-univ.archives-ouvertes.fr/hal-02183031
Submitted on 14 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Organ Segmentation in CT Images With Weak Annotations: A Preliminary Study

Rosana EL JURDI, Caroline PETITJEAN, Paul HONEINE, Fahed ABDALLAH

1LITIS Lab, Université de Rouen Normandie, Saint-Étienne-du-Rouvray, France
2Université Libanaise, Hadath, Beyrouth, Liban
3ICD, M2S, Université de technologie de Troyes, Troyes, France.

rosana.el-jurdi@univ-rouen.fr, caroline.petitjean@univ-rouen.fr
paul.honeine@univ-rouen.fr, fahed.abdallah76@gmail.com

Abstract – Medical image segmentation has unprecedented challenges, compared to natural image segmentation, in particular because of the scarcity of annotated datasets. Of particular interest is the ongoing 2019 SegTHOR competition, which consists in Segmenting THoracic Organs at Risk in CT images. While the fully supervised framework (i.e., pixel-level annotation) is considered in this competition, this paper seeks to push forward the competition to a new paradigm: weakly supervised segmentation, namely training with only bounding boxes that enclose the organs. After a pre-processing step, the proposed method applies the GrabCut algorithm in order to transforms the images into pixel-level annotated ones. And then a deep neural network is trained on the medical images, where several segmentation loss functions are examined. Experiments show the relevance of the proposed method, providing comparable results to the ongoing fully supervised segmentation competition.

1 Introduction

Medical image segmentation is of great importance in medical image computing, at the intersection of several fields in image processing, computer vision, and medicine. It consists in partitioning an image into meaningful segments, such as different tissue classes or distinct organs. While there has been a large effort to address image segmentation of natural images thanks to availability of large annotated databases (e.g., ImageNet with more than 14 millions hand-annotated images, including more that 1 million images bounding box annotations), medical image segmentation is more challenging due to many difficulties. On one hand, medical images encompass segmentation ambiguities, due to low contrast and noise. On the other hand, they are diverse by nature, depending on the region under study and the imaging equipment, such as computed tomography (CT) scanners collecting radiodensity values, and PET scanners for positron emission tomography. For all these reasons, there is no large annotated database that allows to efficiently pre-train or train deep neural networks for medical image segmentation.

In particular, the segmentation of organs in CT images is of great interest. Before radiotherapy, the process of irradiation planning on CT images requires the delineation of the target tumor and healthy organs, called Organs At Risk (OAR), near the target tumor. In practice, the delineation process is manually performed by a medical practitioner. Such a time-consuming approach is often susceptible to an unaffordable level of imprecision, which may result in missed tumorized areas, or attacking a healthy tissue. Therefore, the need for an automated segmentation system has been attracting increasing attention.

In this regard, the Segmentation of THoracic Organs at Risk in CT images (SegTHOR) dataset is of great interest, because each CT image is manually pixel-wise segmented by an expert radiation oncologist [10]. This dataset has just been released publicly in an ongoing competition conducted at the 2019 IEEE International Symposium on Biomedical Imaging.

1https://competitions.codalab.org/competitions/21012
2https://biomedicalimaging.org/2019/challenges/
In this paper, we push forward the segmentation problem, with application to SegTHOR, into the paradigm of weakly supervised segmentation [6]. Our major motivation is the difficulty in obtaining fine-grained pixel-level annotations in medical images. Within the weakly supervised segmentation paradigm, we seek to learn the segmentation by using a training dataset with a rough bounding box annotation that encloses the organ under scrutiny.

Our approach is based on three main stages. The first stage operates a pre-processing in order to transform the SegTHOR CT images into an appropriate format, namely slice/label pairs; it is worth noting that the raw images are very different from natural images (e.g. ImageNet or Pascal VOC). In the second stage, we transform the images with bounding boxes into images having pixel-level annotations. To this end, we use the GrabCut method, which is an iterative intensity graph based method that segments objects from bounding boxes [7]. The third stage is the machine learning (ML). To this end, we consider the FCN (Fully Convolutional Network) and propose to train the neural network on the SegTHOR dataset, by exploring several loss functions that are relevant for image segmentation. A specific attention is carried out on the class imbalance issue, since the four classes are very imbalanced (i.e., organs of different volumes). Note that our method shares similarities with Khoreva et al’s ‘Simple Does It’ method [2]; however different from them, we did not use any objectness cues. Experiments conducted on the SegTHOR dataset demonstrate the relevance of the proposed approach.

The rest of this paper is organized as follows. Next section presents the SegTHOR dataset. Section 3 describes the proposed method of weakly supervised segmentation. Section 4 provides experimental results on the SegTHOR dataset.

2 SegTHOR Dataset Description

The SegTHOR dataset is constituted of 60 patients with lung cancer referred for radiotherapy at the Centre Henri Becquerel, Rouen, France. The CT images have $512 \times 512$ pixels with in-plane resolution varying between 0.90 mm and 1.37 mm per pixel, depending on the patient. The number of slices varies from 150 to 284 with a z-resolution between 2 mm and 3.7 mm. Note that in this paper we focus on the segmentation of one organ at risk, the heart - even though the SegTHOR dataset contains the ground truth segmentation for three other OAR, namely aorta, trachea, esophagus (Fig. 1), that will be considered in upcoming, multiclass studies. For the training, as well as the evaluation, the images were hand-annotated at the pixel level by a radiotherapist. We use the data partition provided in the SegTHOR competition: the training set includes 40 CT scans and the test set includes the remaining 20 CT scans. Note that we extracted the slices in which the organ of interest (e.g. the heart) is present: out of 7390 slices in the provided training set, we kept 1522 of them; in the testing set, 726 images were retained out of 3694 slices. Additionally, we partition the training dataset and leave out four patients for validation (hyperparameter tuning). However, in our case, the training images are not pixel-level segmented, but using bounding boxes that enclose the organs under study. The patient distribution over the folds is shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th># Patients</th>
<th># Slices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training dataset</td>
<td>38</td>
<td>1522</td>
</tr>
<tr>
<td>Validation dataset</td>
<td>2</td>
<td>77</td>
</tr>
<tr>
<td>Evaluation dataset</td>
<td>20</td>
<td>726</td>
</tr>
</tbody>
</table>

3 Proposed Method

3.1 Pre-processing

The pre-processing phase includes clipping pixel intensities into a lower and upper bound of 1000 and 3000. To avoid problems of exploding and vanishing gradients, the data is then normalized by extracting the mean value per image.

3.2 From weakly supervised to supervised data

In this paper, we adopt Grabcut, an iterative intensity based algorithm primarily implemented on RGB images as stated in [7]. Essentially, the algorithm estimates object segments from their bounding box as follows. Given an image and a bounding box(BB), the algorithm considers all pixels outside the BB as belonging to the set $T_B$ of background pixels and considers all pixels within the BB as belonging to $T_U$: the set of pixels to be predicted as foreground or background. The algorithm then estimates two Gaussian mixture models basing on these pixel distributions and builds a similarity graph from the different pixels constituting the image. This graph is later on optimized using minimal cut and new Gaussian distributions are then estimated. Pixels are then reassigned new labels basing on the Gaussian models. The process is iterated until convergence.

3.3 Training the segmentation model

We take advantage of the well-known FCN model [3], a state-of-the-art model for semantic segmentation, in order to address
our medical image segmentation problem. The FCN model resembles an auto encoder, where the encoder is a VGG-16 model with its classification layers removed. Thus, the encoder part is a standard network, which can be pretrained using ImageNet for example. The rest of the network is dedicated to perform image to image inference and can be initialized randomly, then fine-tuned on the SegTHOR data set. Skip-grams are added in order to combine context with spatial information, as suggested in [3].

Taking a look at the nature of the data at hand, we verify the fact that medical images do indeed pose a challenge due to the nature of class imbalance within the medical slices. To solve this problem, we perform a fair comparison of multiple loss functions present within the literature of image segmentation, in order to study their effect on class imbalance. In the following, we provide a brief description of the used loss functions. Let \( p_i \) be the predicted pixel level probability for pixel \( i \), and \( g_i \) its ground truth label. Let \( N \) be the total number of pixels of the image under scrutiny.

When considering binary segmentation, the standard way is the so-called log-loss or cross entropy [5], defined as

\[
- \sum_{i=1}^{N} g_i \log(p_i).
\]

This formulation is sensitive to class imbalance, which is a major issue medical image segmentation because background often overshadows the segmented organs. In the following, we explore loss functions that are less sensitive to class imbalance.

To overcome class imbalance, a weighted version of the cross entropy [1], called balanced cross entropy, is defined as:

\[
- \sum_{i=1}^{N} \beta_i g_i \log(p_i),
\]

where \( \beta_i \) is the ratio of negative samples over total samples. Another metric that takes into account class imbalance is the Dice loss, which has an inbuilt normalization. Dice loss is a soft approximation of the Dice score, defined as the intersection over union ratio of two binary segmentations given a smoothing variable \( \epsilon \). This approximation is required to make the loss function differentiable and turns the binary segmentations into the probabilities: [4]:

\[
\frac{2 \sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \epsilon}.
\]

Similarly as the Dice loss, the Tversky loss [8] is defined as:

\[
\sum_{i=1}^{N} p_i g_i + \alpha \sum_{i=1}^{N} (1 - g_i) + \beta \sum_{i=1}^{N} (1 - p_i) g_i + \epsilon,
\]

where \( \alpha \) and \( \beta \) are tunable parameters. In our experiments we adopt the ratios presented in [8], taking \( \alpha = 0.3 \) and \( \beta = 0.7 \) so that the Tversky loss boils down to the well-known \( F_1 \) score. The latter provides a trade-off between precision (related to false positives) and recall (related to false negatives).

4 Results and Analysis

For evaluation, we have used the Dice coefficient. Our first attempt was to examine the performance of the neural network pre-trained on the ImageNet dataset, without any fine-tuning. Doing this yields a poor segmentation accuracy of 4.67 \% in mean Dice coefficient. This poor performance was expected, because medical image segmentation often requires domain specific knowledge, absent within the features extracted from natural datasets such as ImageNet and Pascal VOC2012. For this reason, we fine-tuned the neural network on the SegTHOR training dataset for heart segmentation.

Despite the fact that image analysis requires minimally a number of few thousand iterations as indicated in [9], however, since our objective is to simply compare the methods as a first step, we settled for about 550 iterations and recorded the results as shown in Tab. 2.

From the table, we realize that balanced cross entropy and binary cross entropy achieved the highest scores with a mean Dice similarity coefficient of 84.67\% and 71.79\%, respectively. In contrast, Dice and Tversky losses scored poorly. This may be due to the nature of these two loss functions, thus requiring a larger number of iterations or larger data-set size to insure stable convergence. Another explanation remains within the poor characterization of the Dice loss function. Thus, the convergence of Dice may highly depend on the value the of smoothing variable \( \epsilon \) set out during training. Thus, smoothing often prevents the gradients from being set to zero. Moreover, it helps the model to overcome the vanishing/exploding gradient problem. On one hand, setting a high value for \( \epsilon \) helps to avoid overfitting, on the other hand a smaller value may enable the training to converge faster.

Performing manual hypertuning of \( \epsilon \), we realize that as the \( \epsilon \) value decreases, the model performance increases. However, with low \( \epsilon \) values, the model is more prone to overfitting. For this reason, proper tradeoff between lower values of \( \epsilon \) and convergence speed should be found. For our experiments, we have adopted a smoothing value of \( 10^{-7} \). Future studies should include proper set up of smoothing value within the Dice, or even a dynamic one that would change with respect to the number of epochs.

After training, we have evaluated the Dice similarity performance with respect to the Grabcut label estimates, in addition to the evaluation w.r.t. the actual ground truth values. We
recorded a similarity value of 94.73 % which at first sight opens way for us to believe that our model is learning well. However, comparing with the actual label masks, we observe that our model registers a performance of 84.67% w.r.t. the ground truth segmentation maps. This causes us to believe that intensity based measures such as grabcuts may not be a good representatives of our dataset labels.

Finally, the proposed weakly supervised method provides an accuracy of up to 84.67% for the heart segmentation, using the balanced cross entropy loss function with 550 iterations, which is comparable to the best results obtained so far at the SegTHOR competition, using fully supervised segmentation methods: the first rank competitor of the leaderboard as of mid-March 2019 has a mean Dice coefficient of 94.32%. However, the score this competitor was obtained on all slices, and not only the ones containing the heart, so his score is not fully comparable. Thus, we can conclude that our weakly supervised approach reaches about 90.86% of the quality of fully supervised models when compared to competition results and about 91.97% when compared to our own fully supervised implementation.

### 5 Conclusion and Future Work

In this paper, we proposed a weakly supervised segmentation method for medical images. Several loss functions were examined. Tested on the heart segmentation of the SegTHOR dataset, the balanced cross entropy loss function outperformed the other losses. With a mean Dice coefficient of 84.67%, it provided comparable results with the best fully supervised methods. In future work, we aim at automatically hypertuning the value $\epsilon$ for proper network convergence under the guidance of the Dice model. We will also integrate our binary segmentation model onto multi-class segmentation. Moreover, we will address the problem of class imbalance by proposing a training procedure that alternates between multiple losses, in order to take into account the strengths of each loss function.

### 6 Acknowledgement

The authors would like to acknowledge the National Council for Scientific Research of Lebanon (CNRS-L) and the Agence Française de la Francophonie (AUF) for granting a doctoral fellowship to Rosana El Jurdi, as well as the ANR (Project API, grant ANR-18-CE23-0014).

### References


