Weakly Supervised Fully Convolutional Network for PET Lesion Segmentation

S. Afshari\textsuperscript{a}, A. BenTaieb\textsuperscript{a}, Z. Mirikharaji\textsuperscript{a}, and G. Hamarneh\textsuperscript{a}

\textsuperscript{a}Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Canada

ABSTRACT

The effort involved in creating accurate ground truth segmentation maps hinders advances in machine learning approaches to tumor delineation in clinical positron emission tomography (PET) scans. To address this challenge, we propose a fully convolutional network (FCN) model to delineate tumor volumes from PET scans automatically while relying on weak annotations in the form of bounding boxes (without delineations) around tumor lesions. To achieve this, we propose a novel loss function that dynamically combines a supervised component, designed to leverage the training bounding boxes, with an unsupervised component, inspired by the Mumford-Shah piecewise constant level-set image segmentation model. The model is trained end-to-end with the proposed differentiable loss function and is validated on a public clinical PET dataset of head and neck tumors. Using only bounding box annotations as supervision, the model achieves competitive results with state-of-the-art supervised and semi-automatic segmentation approaches. Our proposed approach improves the Dice similarity by approximately 30\% and reduces the unsigned distance error by approximately 7 mm compared to a model trained with only bounding boxes (weak supervision). Also, after the post-processing step (morphological operations), our weak supervision approach differs only 7\% in terms of the Dice similarity from the quality of the fully supervised model, for segmentation task.

Keywords: Segmentation, PET scans, Weak annotation, Deep Learning, Fully convolutional network

1. INTRODUCTION

Most of the success of machine learning and deep learning based medical image analysis tasks (e.g. classification, detection or segmentation) so far can be attributed to supervised learning, for which a relatively large annotated set of images must be made available. However, it is well-known that existing annotated medical imaging datasets are scarce and the effort involved in collecting them (especially ones with high quality annotated dense label-fields, i.e. segmentation masks) can be insurmountable.\textsuperscript{1} It is also known that some annotations are easier and faster to collect than others. For instance, localizing lesions with bounding boxes or seeds is less laborious than delineations. In this work, we develop a method to leverage such weak annotations in the context of PET lesion segmentation, where a deep learning model is trained to delineate lesions in 3D PET images from only bounding box annotations around the lesions of interest.

FDG-PET images are used for diagnosis, staging and monitoring cancerous lesions in radiation therapy.\textsuperscript{2} Medical image processing and analysis of PET images is used to detect and localize cancerous lesions in order to quantify tumor metabolic activity from the segmented tumor volume and to design radiation therapy treatment plans. This, in turn, requires more reliable, accurate and reproducible PET image tumor delineation.\textsuperscript{3} An important challenge in designing automatic lesion segmentation techniques in PET images is handling the low signal to noise ratio and the low resolution of these images.

Some segmentation techniques leverage anatomical information from CT scans registered to PET, especially for tumours located near the lungs, where the air creates high contrast in the tumor areas. Otherwise, using CT may cause over-estimations of tumor lesions.\textsuperscript{4} Assuming CT scans are available (which cannot be guaranteed), PET-CT registration is prone to error and, even with hybrid PET/CT scanners, the two modalities are not perfectly aligned because of patient movement or respiratory motion.\textsuperscript{5} Further, certain tumour boundaries are difficult to see in CT due to similarities in Hounsfield units (HU) of tumour lesions and surrounding tissues, so delineating tumours based on metabolic activity presented in PET may be more reliable. In this work, we focus on head and neck lesion segmentation only from PET.
2. RELATED WORK

Different methods have been proposed for delineating lesions in PET images and can be categorized into four major groups. (i) The methodology of the first group is based on selecting either fixed or adaptive thresholds of the maximum standardized uptake values ($SUV_{max}$). However, this approach often performs poorly when applied to real clinical data as lesions generally have irregular shapes, values may vary depending on the reconstruction method used to acquire the PET image, and the images generally suffer from low signal to noise ratios and heterogeneous radio-tracer distributions. (ii) The second group consists of semi-automatic approaches that leverage user input to guide the segmentation algorithm or refine its result. Region of interest (ROI) and seed-guided region-growing methods belong to this group. (iii) The third category are energy-minimization segmentation methods, which are designed to locate and identify the boundaries of the target objects in PET images. Active contour-based models based on the Mumford-Shah functional are examples of this category. These models typically require an initial contour or surface to start the optimization-based segmentation. (iv) The final category encompasses machine learning-based methods, which estimate the segmentation of a novel test image based on the learned statistics of labeled training data. Among the mentioned approaches, machine learning based methods achieved high performance in terms of accuracy and time. One of the key challenges, however, with supervised machine learning techniques (especially deep learning) is their reliance on the availability of a large set of training data.

To address the need of annotation for training machine learning models, different weakly supervised segmentation models that leverage bounding box annotations around objects have been proposed for medical and non-medical segmentation tasks. For natural images, deep convolutional neural network (CNN) models that exploit bounding box annotations give excellent results in semantic segmentation tasks. For medical image segmentation, few works use CNN models to leverage bounding box annotation. One of the closest works to ours is Rajch et al.’s, where the authors built a model similar to the popular GrabCut algorithm by replacing the Gaussian mixture model (GMM) with a CNN whose network parameters are optimized by minimizing a conditional random field (CRF) energy function. They applied their method to segment the brain and the lungs from MR images using bounding box annotations. While our proposed approach aims at leveraging a similar type of weakly supervised training data for medical image segmentation, our work has important differences: We are segmenting lesions in PET (not organs from MR); our method is truly 3D whereas other works process 3D volumes as 2D slices, which results in loss of context when segmenting 3D lesions; and our approach does not require user interaction at inference time whereas other methods do.

Another closely related work to ours is Deep Level Sets by Hu et al. At a high level, both our method and theirs combine a convolutional network with energy-minimization based segmentation. In contrast, while our model is based on the 3D network architecture, we formulate a different level-set energy as a regularization term in a new loss function to segment target lesions in PET images in an unsupervised manner. This regularization term allows us to train the model in a weakly supervised context (i.e. using only bounding box annotations for training).

3. METHOD

3.1 Novel Loss Function

To leverage weakly supervised data yet achieve 3D lesion delineation at test time without any user-interaction, we propose a novel loss function that dynamically combines two loss terms; one that is guided by the training bounding boxes, and the other designed to realize Mumford-Shah-inspired piecewise constant segmentation within the box. We show competitive results with other works on a challenging public dataset of PET lesions and compare our model to different baseline methods, experts’ delineations, and other semi-automatic approaches.

Given a set of 3D PET training images and their corresponding bounding box annotations, our goal is to optimize the parameters of an FCN in an end-to-end manner to identify and segment all lesions of interest. Our model is trained with a new loss function:

\[ L = \beta L_{MDice} + \lambda L_{MS} \]
where \( L_{MDice} \) is the modified version of the Dice loss function, \( L_{MS} \) is the Mumford-Shah inspired term and \( \beta \) and \( \lambda \) are hyper-parameters balancing the contribution of each term. The interaction between both terms of the loss function is controlled via these hyper-parameters and allows the network to first predict segmentations that are relatively close to the real segmentation masks, and then refine them using the regularization term.

In segmentation tasks where the majority of voxels are background voxels, the Dice similarity coefficient is often deployed as a loss function, \( L_{Dice} \), to address this class imbalance:

\[
L_{Dice} = 1 - \frac{2 \sum_{i=1}^{N} p_i \cdot b_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} b_i}
\]

where \( p_i \) and \( b_i \), in order, are the prediction label and the ground truth label for a voxel \( i \), and \( N \) is the total number of voxels in the volume.

To encode the Mumford-Shah piecewise constant image model, commonly used with energy minimizing level-set segmentation, and train a deep network under weak supervision, we modify \( L_{Dice} \) as follows.

### 3.2 Modified Dice loss term

Optimizing \( L_{Dice} \) when annotations are bounding boxes forces the network to generate box-like masks, which is undesirable for a lesion segmentation task. To generate segmentation masks instead of bounding boxes, we modify the denominator of \( L_{Dice} \) by proposing a new term \( L_{MDice} \) that penalizes the misclassified voxels inside and outside of the box annotations in different ways. Outside the box, the classical Dice loss is always applied (to discourage any false positive voxels). Inside the box, however, the loss is dynamically modified such that the Mumford-Shah piecewise constant image model takes effect, instead of the classical Dice, whenever sufficient voxels are labeled as foreground. This dynamic loss behavior trains the network to use the weakly supervised labeling to effectively localize lesions and the Mumford-Shah model to delineate them. In particular, we set:

\[
L_{MDice} = 1 - \frac{2 \sum_{i=1}^{N} p_i \cdot b_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} b_i - M \cdot H(\frac{\sum_{i=1}^{N} p_i \cdot b_i}{\sum_{i=1}^{N} b_i})}
\]

where \( H(x) \) is a Heaviside function shifted by hyper-parameter \( \alpha \). Note that \( L_{MDice} \) is equal to \( L_{Dice} \) as long as the fraction of voxels, inside the ground truth bounding box and classified as foreground, is less than or equal to \( \alpha \). Otherwise, the denominator of \( L_{MDice} \) becomes \( \sum_{i=1}^{N} p_i + \sum_{i=1}^{N} p_i \cdot b_i \), which means that voxels inside the box yet predicted as background are not penalized; this is when \( L_{MS} \), described next, is most critical.

### 3.3 Mumford-Shah inspired loss term

In order to train our network to produce segmentation masks, inside the bounding box, that mimic results obtained using energy-minimizing level-set methods that encode the Mumford-Shah piecewise constant image model (e.g.,\(^{13} \)), we define \( L_{MS} \) as follows:

\[
L_{MS} = \frac{1}{\sum_{i} b_i} \left( \omega_1 \sum_{i} (|I_i \cdot H_\epsilon(\phi_i) - C_1| - |I_i \cdot H_\epsilon(\phi_i) - C_2|) \cdot b_i + \omega_2 \sum_{i} (|I_i \cdot (1 - H_\epsilon(\phi_i)) - C_2| - |I_i \cdot (1 - H_\epsilon(\phi_i)) - C_1|) \cdot b_i \right)
\]

where \( I_i \) is the input image intensity of the \( i^{th} \) voxel. \( H_\epsilon(x) \) is the regularized Heaviside function proposed in:\(^{13} \)

\[
H_\epsilon(x) = \frac{1}{2} \left( 1 + \frac{2}{\pi} \arctan\left( \frac{x}{\epsilon} \right) \right).
\]
We set $\epsilon$ to 0.03 as suggested in. $\phi$ is the level-set function, which we obtain by linearly shifting the prediction map values to be in the range $[-0.5, 0.5]$, so as to encode the lesion boundary as its zero level-set. Multiplying by $b_i$ (defined in (2)) confines $L_{MS}$ calculations to the interior of the bounding box. Normalization by $\sum_i b_i$ is used to obtain similar contribution to the loss from bounding boxes of different sizes. $\omega_i$ are scalar weights, and $C_1$ and $C_2$, in order, are the average intensities of voxels inside and outside of the prediction boundary, which are calculated by:

$$\quad C_1 = \frac{\sum_i I_i \cdot H(\phi_i) \cdot b_i}{\sum_i H(\phi_i) \cdot b_i} \quad \text{and} \quad C_2 = \frac{\sum_i I_i \cdot (1 - H(\phi_i)) \cdot b_i}{\sum_i (1 - H(\phi_i)) \cdot b_i}$$ (7)

### 3.4 Neural Network Architecture

We adopt the U-Net architecture with contracting path to efficiently exploit contextual information. U-Net demonstrated good performance when the number of training images is limited. We changed the 2D architecture of U-Net to 3D FCN and reduced the total number of feature maps by a factor of 4. The size of the input image, the output segmentation mask, and the ground truth mask is set to 160 $\times$ 160 $\times$ 128. Scaled exponential linear units (Selu) are selected as the activation functions in the middle layers and sigmoid function for the output layer. The model is trained end-to-end with the proposed loss function in 1.

### 4. EXPERIMENTS

#### 4.1 Image Data

We applied our method to 57 FDG-PET scans of unique patients from the public collection of head and neck cancer provided by the Quantitative Imaging Network of the US National Cancer Institute. From this dataset, a subset of 10 PET scans was utilized for the QIN PET segmentation challenge. On average, the dataset includes 3.8 ± 2.5 lesions per PET volume (ranging from 1 to 12) with the volume ranges from 1 to 13 m$^3$. The selected test cases included different complexity levels (ranging from low to high). The PET images vary in resolution from 128 $\times$ 128 to 168 $\times$ 168 and the number of axial slices ranged from 191 to 545. The average voxel size in this dataset is 3.5 $\times$ 3.5 $\times$ 3.1 mm.

We used the challenge test set and the provided delineation by experts (average over experts’ delineations to reduce inter- and intra- variability) for evaluation of our model and all other baselines reported in the results. We train the network on the remaining 47 cases. The box annotations are coarse bounding boxes around all lesions for each image with an average Dice similarity, with the ground truth delineations, of 28% ± 11%.

Given the low resolution of PET images and the complexity of the task, we found the following pre-processing steps necessary to accurately train our model and all baselines. First, as this dataset only includes lesions in the head and neck area, we reduced the computational cost by cropping all volumes inferior to the neck area such that we could process the entire volumes (not patch-based, and not slices) with 3D networks. We normalized all images to zero mean and unit variance and performed different data augmentation strategies (i.e., random flipping left-right and superior-inferior orientation, and additional Gaussian noise). After augmentation, the total number of images in the training and validation sets were 430 and 40, respectively.

#### 4.2 Comparative Quantitative Evaluation

We compare our method with a fixed architecture CNN trained with fully supervised ground truth delineations by experts ($L_{FS}$) and a CNN trained with bounding box annotation ($L_{Dice}$). The performance of these two networks can be considered as the upper bound and the worst-case accuracy, respectively, that our method can achieve. In addition, we include the result of training our model with only $L_{MDice}$. We also tested using morphological operations (dilation and erosion) as a post-processing ($L_{MDice} + L_{MS} + M$) to show that our result can be improved by separating boundaries of touching lesions and removing small sized false positive predictions. Furthermore, we compare our results to those obtained when active contours without edges level-sets method is applied, as a post-processing step, to the prediction output of $L_{Dice}$ ($L_{Dice} + LS$). We also added the result of the semi-automatic approach ($SA$), which is proposed for segmenting tumor lesions in head and neck from PET.
images. To compare with experts performance, we include the experts agreement by averaging Dice coefficient over all experts within each trial (Experts).

4.3 Setup and Parameter Selection
For training the network, all hyper-parameters are tuned on the validation set. For the proposed loss function, \(\omega_1\) and \(\omega_2\) are selected empirically based on the strength of two terms in equation 5 for the validation dataset, and are set to 1 and 3, respectively. \(\beta\) and \(\lambda\) are set to 10 and 1, respectively, in order to emphasize the Dice loss. For the modified Dice loss term, we used \(\alpha = 0.1\), which means at least 10% of the voxels inside the box are forced to be predicted as tumor. We used stochastic gradient descent as the model optimizer and set the momentum, weight decay, and learning rate to 0.9, \(1e^{-6}\) and \(1e^{-5}\), respectively. The batch size is 1 and the maximum number of epochs is 100, which is limited by an early stopping criteria when there is no improvement on validation loss after 20 epochs. We used the trained network on box mask annotations to initialize the weights.

4.4 Results
To evaluate the performance of the proposed framework, we adopted the Dice similarity coefficient and unsigned distance error (UDE) metrics, used in.\(^{16}\) Figure 1 shows that the improvement of the Dice coefficient is approximately 30% greater than the performance of the same model trained using the original Dice \(L_{\text{Dice}}\). We also found that by applying morphological operations to our weakly supervised outputs, our performance of \(L_{\text{MDice}} + L_{\text{MS}} + M\) differs from the results of the fully supervised \((L_{FS})\) approach trained by the expert’s delineations by only 7% in Dice similarity. Furthermore, while our method does not need either user interaction or ground truth segmentation masks, its performance is not very far from the performance of the \(SA\) approach, in which a user selects a point in the center of each lesion of interest (Dice similarity 68% and 82%, respectively). The mean UDE for the proposed method is 3.1 mm, which is comparable to the \(L_{FS}\) model with an error of 2 mm. Figure 2 shows examples of segmentation results for two different 3D PET images. For the first row, the proposed method is successful in delineating the tumor lesions while in the second image, it fails to find and delineate all lesions.

![Figure 1. Mean Dice and Mean unsigned distance error (UDE) for different methods. The error bars represent the standard deviation of the metrics for each method.](image)

5. CONCLUSION
Given the clinical importance of segmenting lesions from PET images (e.g., for radiation therapy treatment planning or assessing treatment efficacy) and given the tedious task of manually delineating datasets for training machine learning algorithms, we proposed a deep learning lesions segmentation method that requires only weakly labeled data (in the form of bounding boxes). We developed a novel multi-loss function with an adaptive Dice
Figure 2. 2D (coronal and sagittal views) and 3D rendering of segmented lesions in two test cases. Green and purple colors, in order, show the lesions segmented by the proposed method and the average of ground truth lesions segmented by experts. In the first row, the proposed method successfully segments the lesions (Dice similarity of 83%). In the second row, our method fails to segment the lesions (Dice similarity of 31%).

coefficient term and a Mumford-Shah piecewise constant model term. Our results validated that the provision of training data in the form of bounding boxes is a viable option to careful expert delineation. Future work will focus on testing the method on other types of cancer lesions.

ACKNOWLEDGMENTS

This research is supported by the Canadian Institutes of Health Research (OQI-137993). Also, we thank the NVIDIA Corporation for the donation of a Titan X GPU used in this work.

REFERENCES