Stereology as Weak Supervision for Medical Image Segmentation

Giorgia Silvestri
Orobix Srl
Bergamo, Italy
giorgia.silvestri@orobix.com

Luca Antiga
Orobix Srl
Bergamo, Italy
luca.antiga@orobix.com

Abstract

Building large medical imaging datasets for image segmentation is a challenging task due to manual outlining. In this work, we explore the use of stereology to cut the costs of annotation. We train a segmentation model using a coarse point counting grid as the sole annotation and quantify the impact of this approach on segmentation performance. Results show that dense masks are not a strict requirement for training segmentation models to achieve satisfying performance. Since deciding whether a small set of grid points overlaps a structure of interest is an inherently faster operation than tracing a dense outline, this method allows to scale up volume annotation to large datasets.

1 Introduction

Segmentation of organs and other substructures from medical images allows quantitative evaluation of morphological parameters related to volume and shape, as, for example, in cardiac, brain analysis or abdominal organs [1]. The most frequently used CNN architecture for medical image segmentation is U-net, published by Ronneberger et al [2], and its derivatives. The characterizing feature of the U-net architecture is the use of skip connections between the up-sampling and down-sampling paths. During training, network parameters are optimized by minimizing a loss function evaluated on the overlap between the network output and a ground truth image. The ground truth is usually provided by human experts as a binary volume representing dense segmentation of the target structure, and as such it requires a considerable amount of time and effort.

Starting from the 1960s, enabled by technological advances in microscopy, the need to obtain 3D information about biological objects based on their 2D appearance arose, giving rise to stereology [3]. The average volume density volume of a target structure can be estimated accurately from the sum of areas on planar sections distanced by T. Such areas can be in turn estimated using a test grid consisting of a set of points p arranged in a rectangular or hexagonal lattice, each with an associated surrounding area a/p. By counting the number of test grid points P falling into the structure of interest, the volume of the structure is estimated as $V = T * a/p * \sum P$.

In this work, we adopt the point counting method as weak supervision of an image segmentation task and demonstrate that training a U-net with a point counting mask incurs in limited decrease in segmentation performance, which is out-weighted by the opportunity of annotating volumes at a much lower cost. Since deciding whether a small set of points on a coarse grid overlaps a structure of interest is an inherently faster operation than tracing a dense outline, this method allows to scale up volume annotation to large datasets. Results also hint at the fact that a higher number of weakly supervised volumes is preferable to a lower number of densely annotated volumes.

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

2 Methods & Experiments

In this work we train a 3D U-Net [2] using both a dense segmentation target as well as targets generated from grids at different point densities on the same data. In the experiments, we focus on hexagonal grids, to maximize the chances of overlap with rounded biological shapes. We denote the horizontal distance between two adjacent grid points as **stride**. Grids are superimposed on each segmented slice and their offset is randomized per slice. Grid points are then manually classified whether or not they overlap with the structure of interest. At each training iteration, we dynamically generate a dense ground truth by padding each grid point on its 2D plane by a variable amount. In one set of experiments we keep the padding amount fixed to $(stride-1)/2$ (**FixP**), while in a second set of experiments we randomize the padding amount on each 2D slice in the range $[0;(stride-1)/2]$ (**RP**). Note that this second case leads to a data augmentation effect, since the padding amount changes independently for each slice at each training iteration. We evaluate our method using point grids of stride $3, 5, 7, 9, 11, 13, 15, 17, 19, 21$ and compare each with the performance obtained using the dense ground truth.

The data employed for this study are 399 high-resolution abdominal CT scans acquired with a slice thickness of $3mm$. A dense segmentation of the pancreas was manually provided by two expert radiologists. The original image size is $512 \times 512 \times N_{slices}$. Scans are clipped between -150 and 500 Hounsfield units and a down-sampling with average and max pooling with factor 2 are applied to image and ground truth, respectively. Each scan is then re-scaled to a $[0, 1]$ range. A random selection of 79 scans are kept for performance comparisons, while the remaining 320 scans are employed for training. Out of this set, a random subset of 100 scans serves a reduced dataset for evaluating the effect of sample size on segmentation performance.

The input to the model are 3D slabs of size $256 \times 256 \times 10$. The z location of the slab is randomized during training. The model has 5 3D convolutional blocks followed by max pooling and $tanh$ activation modules on both the up-samling and down-sampling path. Batch normalization is applied to the input of the network. We train the model using the Adam optimizer. The learning rate starts at $1e^{-4}$ and it is reduced by a factor of 0.5 if no improvement is seen for 200 epochs. We use the complement of the *Dice coefficient* as loss function. The Dice coefficient D between two binary volumes ranges between 0 and 1 and can be written as $D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}$, where the sum runs over N voxels of the predicted binary segmentation volume $p_i \in P$ and the mask binary volume $g_i \in G$, where P is network output and G is the grid generated mask. During training, the model aims to minimize $1 - D$. During every training iteration, we feed randomly rotated versions of the training images as input. The applied rotation is a random rotation around the z axis in the range $[-30, 30]$ degrees. All code is written in Python, using the PyTorch deep learning library (`http://pytorch.org`).

3 Results

During inference, the scan is subdivided into consecutive slabs spaced by one slice. An average of all predictions at each voxel is assumed as the output, no further post-processing is applied. Results are reported in Table 1 and in Figure 1 and show that weak supervision achieves is within a few percent accuracy from full supervision for all grids. Performance obtained using a grid of stride 5 on 320 scans exceed the ones obtained using a dense target on 100 scans. Using large strides (greater than 9) randomizing padding is a key factor for achieving good segmentation performance.

References

[1] Geert Litjens, Thijs Kooi, and et al. Bejnordi. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[3] Marshall D Sundberg. An introduction to stereological analysis: morphometric techniques for beginning biologists. *Chapter*, 3:70803–1705, 1992.

Table 1: Results for all experiments. Volume density is a ratio between the number of grid points overlapped with the ground truth and the number of voxels of the ground truth. All performances are evaluated on 79 independent scans and expressed in terms of Dice coefficient (standard deviation).

Stride grid	Volume density (%)	100 scans (RP)	320 scans (FixP)	320 scans (RP)
ground truth	100	0.78 (0.12)	**0.85** (**0.09**)	**0.85** (**0.09**)
3	11.11 (0,10)	0.73 (0.14)	**0.82** (**0.09**)	**0.81** (**0.12**)
5	4.00 (0,07)	0.71 (0.15)	**0.81** (**0.09**)	**0.80** (**0.12**)
7	2.04 (0,06)	0.71 (0.14)	0.79 (0.10)	0.77 (0.13)
9	1.23 (0,05)	0.68 (0.16)	0.78 (0.09)	**0.80** (**0.09**)
11	0.82 (0,05)	0.71 (0.12)	0.76 (0.10)	0.79 (0.10)
13	0.59 (0,04)	0.72 (0.11)	0.73 (0.09)	0.77 (0.10)
15	0.45 (0,04)	0.71 (0.11)	0.70 (0.10)	0.77 (0.10)
17	0.35 (0,04)	0.68 (0.14)	0.67 (0.11)	0.73 (0.12)
19	0.28 (0,03)	0.66 (0.13)	0.65 (0.09)	0.74 (0.10)
21	0.23 (0,03)	0.58 (0.17)	0.64 (0.10)	0.75 (0.11)

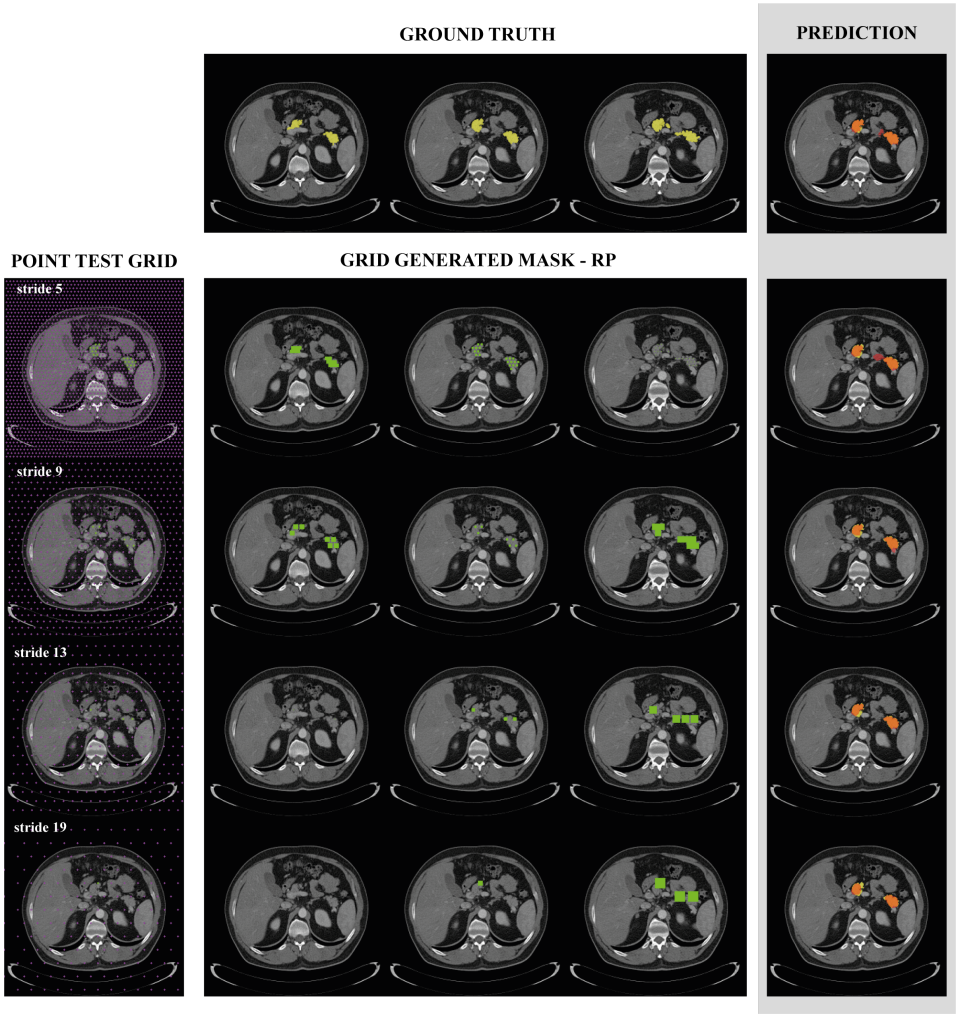

Figure 1: First row: ground truth (yellow) superimposed on 3 CT slices and the prediction (red) on the mid CT slice. From second to fifth rows: test grid with different stride superimposed on the mid CT slice (crosses on and off the ground truth are in green and pink, respectively), examples of padded grids and the prediction (red) on the mid CT slice.