Learning from sparsely annotated data for semantic segmentation in histopathology images

John-Melle Bokhorst[1] JOHN-MELLE.BOKHORST@RADBOUDUMC.NL
Hans Pinckaers[1] HANS.PINCKAERS@RADBOUDUMC.NL
Peter van Zwam[2] P.VANZWAM@PAMM.NL
Iris Nagtegaal[1] IRIS.NAGTEGAAL@RADBOUDUMC.NL
Jeroen van der Laak[1] JEROEN.VANDERLAAK@RADBOUDUMC.NL
Francesco Ciompi[1] FRANCESCO.CIOMPI@RADBOUDUMC.NL
[1] *DIAG Nijmegen, Geert Grooteplein Zuid 10, 6525 GA The Netherlands*
[2] *PAMM Laboratory for Pathology and Medical Microbiology, Eindhoven, The Netherland*

Abstract

We investigate the problem of building convolutional networks for semantic segmentation in histopathology images when weak supervision in the form of sparse manual annotations is provided in the training set. We propose to address this problem by modifying the loss function in order to balance the contribution of each pixel of the input data. We introduce and compare two approaches of loss balancing when sparse annotations are provided, namely (1) instance based balancing and (2) mini-batch based balancing. We also consider a scenario of full supervision in the form of dense annotations, and compare the performance of using either sparse or dense annotations with the proposed balancing schemes. Finally, we show that using a bulk of sparse annotations and a small fraction of dense annotations allows to achieve performance comparable to full supervision.

Keywords: Weakly supervised semantic segmentation, loss balancing, partially labelled data, computational pathology.

1. Introduction

The ability of computers to extract information from images has increased tremendously since convolutional neural networks (CNNs) have been introduced. For multiple years now, CNNs have been successfully applied to classification and segmentation tasks. Segmentation in medical imaging is the process of delineating the boundaries of various structures or tissues. As an example, in histopathology images of colorectal cancer (CRC), distinguishing glands (both healthy and cancerous) from surrounding connecting tissue (i.e., stroma) can be the basis of prognostic biomarkers, such as the tumor-stroma ratio (Mesker et al., 2007) (Geessink et al., 2019).

In semantic segmentation, supervised training of models usually requires labor intensive pixel annotations, which consist in a *dense* segmentation map (Figure 1(a)). In this approach, all pixels, mostly within a pre-fixed area, are assigned to one class by a human annotator. In the field of medical imaging and in particular of histolopathology, this approach is not only labor intensive but also requires specialist knowledge about the transition between the different tissue types. Dense annotations allow a model to learn the transition between different classes, which is expected to produce an accurate semantic segmentation output. This approach can be considered as full supervision of segmentation models.

© 2019 J.-M. Bokhorst, H. Pinckaers, P. van Zwam, I. Nagtegaal, J. van der Laak & F. Ciompi.

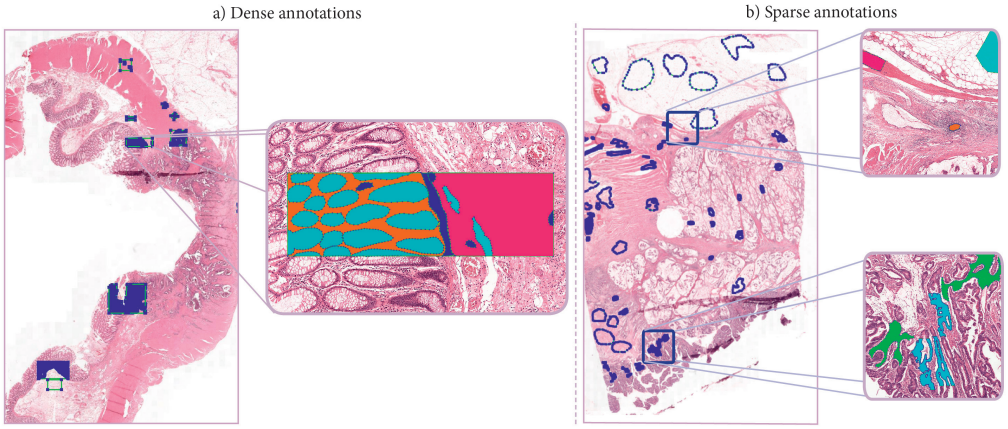

Figure 1: a) Example of a densely annotated image, b) Example of a sparsely annotated image.

An alternative to a fully supervised approach to segmentation is weak supervision, which can be provided in the form of bounding boxes, image level labels, dots or *sparse* (or partial) annotations (Figure 1(b)). In all these cases, only one or more parts of a tissue/class is labeled. In (Rajchl et al., 2017), bounding boxes were used for the segmentation of brain tissue, while (Kervadec et al., 2018) adapts the loss function for segmenting cardiac images with data annotated with scribbles. In (Glocker et al., 2013), a semi-automatic labeling strategy is proposed where sparse dot annotations are converted into dense probabilistic labels for vertebrae localization and naming. In histopathology images, (Xu et al., 2014) proposed to segment glands in CRC based on bounding boxes, however this method is not based on convolutional networks.

Sparse annotations allow to easily include more pixels than a scribble-based approach, but on the one hand do not guarantee to provide clear definition of transitions between different classes, and on the other hand provide pixel-level labels without the localization carried by bounding boxes. One typical use case of sparse annotations is focusing only on areas where the expert is absolutely certain about a specific class. Furthermore, it allows to quickly annotate a large variety of tissue types without having to focus on the surrounding of each specific class, which helps in the case of semantically under-represented tissues. For these reasons, sparse annotations may be considered as an attractive approach to create reference standard in medical imaging for building supervised segmentation models.

In this paper, we address the problem of multi-class semantic segmentation when *sparse* annotations are provided. For this purpose, we tackle the problem of class imbalance and lack of annotated pixels in training examples by modifying the loss function. We formulate two strategies to weigh the loss, namely (a) *instance based balancing* and (b) *mini-batch based balancing*. To investigate the proposed approach, we also consider a set of dense annotations and train segmentation models on fully annotated images alone as well as models that are given mainly sparsely annotated images and a few densely annotated images. We validate our approach on a tissue segmentation problem in colorectal cancer histopathology images, and we use U-Net as the CNN architecture for semantic segmentation. To the best of our knowledge we are the first to try semantic segmentation with sparsely annotated data in CRC histopathology images.

2. Materials

Seventy paraffin-embedded tissue specimens from colorectal cancer patients of the Radboud University Medical Center (Nijmegen, Netherlands) were included. Tissue slides were prepared and stained with H&E staining, and digitalized using a Pannoramic P250 Flash II scanner (3D-Histech, Hungary) at a spatial resolution of 0.24 $\mu m/$px.

A pathologist and two trained human analysts were involved in manual annotations of whole-slide images. The set of cases was split into two parts and both sparse and dense annotations were made: sparse annotations were made on 54 images, dense annotations were made on 16 images. In order to make dense annotations, areas of different sizes, showing at least 2 tissue classes and the border area between them, were selected and annotated. In all images the following 13 tissue types where annotated; 1) tumor, 2) desmoplastic stroma, 3) necrosis and debris, 4) lymphocytes, 5) erythrocytes, 6) muscle, 7) healthy stroma, 8) fatty tissue, 9) mucus, 10) nerve, 11) stroma lamina propria, 12) healthy glands, 13) background. The ratios between the amount of annotated pixels per class for both datasets is shown in Table 1.

	Dense	Sparse
Tumor	17.43	4.47
Desmoplastic stroma	13.68	5.16
Necrosis and debris	1.36	9.80
Lymphocytes	1.80	3.83
Erytrocytes	0.67	2.34
Muscle	13.53	29.84
Healthy stroma	15.98	10.96
Fatty tissue	6.74	24.00
Mucus	7.73	5.86
Nerve	0.18	0.40
Stroma lamina propria	7.21	0.55
Healthy glands	6.35	0.75
Background	7.33	2.02

Table 1: Percentage of pixels per class in datasets annotated with sparse and dense annotations.

The set of whole-slide images (WSI's) with corresponding annotations was randomly divided into a training set (43 WSI's with sparse annotations, 8 WSI's with dense annotations), a validation set (11 WSI's with sparse annotations, 2 WSI's with dense annotations) and a test set, containing 5 WSI's with only dense annotations.

3. Method

When training a segmentation network like U-Net with mini-batch gradient descent (i.e., mini-batch size > 1), attention should be paid to the contribution of individual pixels to the loss function. When sparse annotations are used it may occur that (1) not all classes are present equally in a mini-batch or within a patch and (2) not all pixels within the patch have been assigned to a label, as shown in Figure 2a. In order to tackle these problems, we investigate the effect of modifying the loss function based on the type of manual annotations of input training data. Inspired by the original work on the U-Net model, we define a weight map W that specifies the contribution of each pixel to the loss function L. In practice, if w_{ij} and l_{ij} are the weight map and the loss value for a pixel in position (i, j), using a the weight map produces a new $\hat{l}_{ij} = w_{ij} l_{ij}$ loss for each pixel. We introduce and compare two strategies to create such a weight map, based on different loss balancing strategies, namely (1) *instance based balancing*, and (2) *mini-batch based balancing*. We also compare these approaches with a case without balancing . These three approaches are formulated in detail in this section.

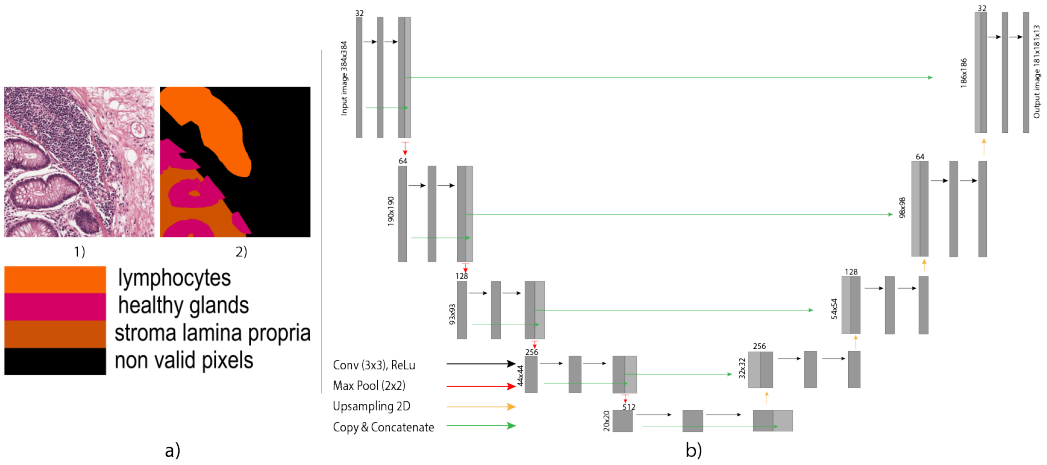

Figure 2: a1) patch from the sparsely annotated dataset with multiple classes, a2) the corresponding annotation map. b) used U-net architecture. Each dark-gray box represents a multi-channel feature map. The input size is shown on the left of the box, and the number of channels on top. The light-gray boxes represent copies of the feature maps. The arrows denote the different operations.

Mask of valid pixels, no balancing. We assume that pixels that have not been annotated in the training set should not contribute to the optimization of the network during training. For this purpose, we define a "mask of valid pixels". In practice, this mask consists of a weight map W of coefficients $w_{ij} \in \{0,1\}$, where $w_{ij} = 0$ is used for pixels that are not annotated (label $y_{ij} = 0$), and $w_{ij} = 1$ is used for pixels that are annotated. In this way, all annotated pixels are considered as "valid" and equally contribute to the loss, not taking into account for a possible class imbalance. We apply this mask to all experiments in this paper, and we also refer to it as a case in which *no balancing* is applied.

Instance based balancing. Both in the case of sparse and dense annotations, a single instance can contain multiple classes with a different amount of pixels per class. Let us define as L_I the amount of valid pixels in an instance (i.e. a training patch) and as C_I the amount of classes present in that instance. To compensate for class imbalance in a patch, we formulated a weight map that ensures that (1) only valid pixels are considered, and (2) all classes contribute the same to the loss:

$$w_{ij} = \begin{cases} \frac{L_I}{C_I C_{ij}}, & \text{if } y_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where C_{ij} represents the amount of pixels belonging to the class in position (i, j).

Mini-batch based balancing. When mini-batch gradient descent is used, an instance based balancing strategy does not take into account the distribution of labels within the mini-batch. In some cases, this may result in some classes having little contribution to parameters update, for example when they only appear in one instance, while other classes may appear in multiple instances of the

	Sparse			Dense			Combined		
	w/o	inst	mb	w/o	inst	mb	w/o	inst	mb
Background	0.32	0.21	**0.34**	0.25	0.25	0.22	0.24	0.25	0.23
Desmoplastic stroma	**0.69**	0.68	0.67	**0.69**	**0.69**	**0.69**	0.65	0.63	0.67
Erythrocytes	0.53	0.49	0.62	**0.72**	**0.72**	**0.72**	0.50	0.68	0.65
Fat	0.84	0.77	0.84	0.85	0.84	0.85	0.80	**0.86**	**0.86**
Healthy glands	0.83	0.86	0.86	**0.88**	**0.88**	0.87	0.84	0.86	0.86
Healthy stroma	0.62	**0.67**	0.66	0.47	0.48	0.48	0.50	0.64	0.59
Lymphocytes	0.82	**0.83**	**0.83**	0.71	0.71	0.72	0.79	0.81	0.76
Mucus	0.22	0.20	0.47	**0.96**	**0.96**	**0.96**	0.42	0.76	0.89
Muscle	0.66	0.61	0.54	0.65	0.65	0.66	**0.70**	**0.70**	0.65
Necrosis and Debris	0.28	0.38	0.39	0.41	0.41	0.41	**0.42**	**0.42**	0.39
Nerve	**0.69**	0.64	0.56	0.62	0.61	0.62	0.55	0.62	0.56
Stroma lamina propria	0.76	0.78	0.76	**0.82**	**0.82**	**0.82**	0.77	0.81	0.79
Tumor	**0.86**	**0.86**	**0.86**	**0.86**	0.85	**0.86**	0.84	0.85	0.84
Overall	0.61	0.62	0.65	**0.68**	**0.68**	**0.68**	0.62	**0.68**	0.67

Table 2: Dice scores for every class per annotation type; w/o refers to no balancing, *inst* to instance based balancing and *mb* to mini-batch balancing.

same mini-batch. For this reason, we extend the concept of balancing to the mini-batch by defining the amount of valid pixels in a mini-batch as L_B and the amount of classes in a mini-batch as C_B. As done for the instance based balancing, each pixel in position (i, j) contributes to the loss with a coefficient w_{ij} computed as follows:

$$w_{ij} = \begin{cases} \frac{L_B}{C_B C_{ij}}, & \text{if } y_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where C_{ij} represents the amount of pixels belonging to the class in position (i, j).

Training. A five level deep U-Net has been chosen as the segmentation network (see Figure 2b). The network architecture is based on the original U-Net paper (Ronneberger et al., 2015) where the number of filters is doubled after every max-pooling layer and the initial filter size is set to 32. Additionally, skip connections within convolutional layers have been added, where the input of the layer block is concatenated with the last feature map. Transposed convolutions have been replaced with up-sampling operations followed by a convolution in the expansion part.

Multiple U-Net models were trained using sparse annotations, dense annotations and a combination consisting of sparsely annotated images and densely annotated images in a ratio of 4:1. The input of all network configurations was a RGB patch of 384×384px with a pixel size of $1\mu m$. For all annotation types all the proposed weight balancing methods were applied.

During training, data was augmented by random flipping, rotation, elastic deformation, blurring, brightness (random gamma), color and contrast changes. An adaptive learning rate scheme was

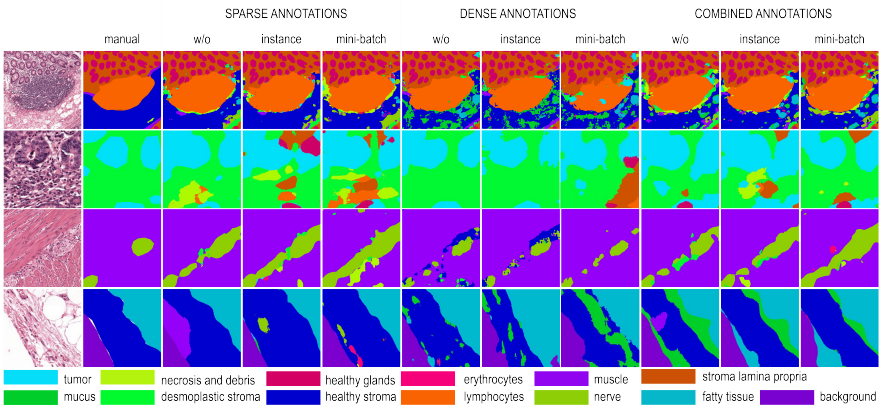

Figure 3: Segmentation output for the considered approaches.

used, where the learning rate was initially set to 0.00001 and then multiplied by a factor of 0.5 after every 10 epoch if no increase in performance was observed on the validation set. The weights of the network were initialized as proposed in (He et al., 2015). The mini-batch size was set to 8 instances per batch, the networks were trained for a maximum of 300 epochs, with 750 iterations per epoch. Categorical cross entropy was used as loss function. The output of all networks is in the form of C likelihood maps. To obtain a final segmentation output the arg-max was taken as the final label.

4. Results

The test set only contained densely annotated regions. From the 5 WSI's used for testing a total of 49 manually annotated regions were selected with an minimum area of 0.375 mm^2 and a maximum area of 0.780 mm^2 per region. From these regions 1250 non overlapping tiles were extracted and segmented by the network. The Dice score was used as performance metric. Dice was calculated for every individual class and as a (class) overall score (see Table 2).

Models trained with dense annotations achieved the best performance (Dice = 0.68). The differences across various balancing methods are marginal without a clear preference for any of the balancing methods.

It can be noted that when sparse annotations were used, mini-batch based balancing outperformed instance based balancing slightly. The instance based balance method gives a slight improvement over training without balancing with Dice scores of 0.62 over 0.61 respectively. Applying mini-batch based balancing shows a better added value with a Dice of 0.65.

A similar trend is observed when sparse and dense annotations are combined. In this case, using instance based normalization allows to achieve a Dice = 0.68, which is comparable to what has been obtained with dense annotations. It is worth noting that comparable performance has been achieved with a significantly reduced amount of dense annotations, namely only 20% of dense annotations.

Visual examples of results for the considered approaches and weight balancing strategies are depicted in Figure 3.

5. Discussion and conclusions

We have introduced different strategies to modify the loss function in semantic segmentation using a U-Net architecture, in order to address the problem of class imbalance and lack of annotated pixels in training examples, namely (a) instance based balancing and (b) mini batch based balancing. The results show that, in training with sparsely annotated images, only considering valid pixels without introducing any balancing strategies gives the lowest performance. Instance weight balancing slightly improves performance, but this annotation method seems to be best supported by weight balancing at the level of the mini batches. This corroborates the validity of a mini-batch based balancing in cases where for example one single class is only present in an instance, which may be penalized depending on the rest of the instances in the mini-batch.

We experimentally observed that the results for dense annotations are not influenced by any balancing strategy. This can be due to the fact that when all pixels are valid and multiple classes are present in an instance, very little variation is caused to the loss when different strategies are used.

When combined annotations are used, the best result is obtained when instance based balancing is applied. This is in contrast with using only sparse annotations, and can be explained by the fact that balancing at mini-batch level in the presence of a few densely annotated instances in the mini-batch eventually penalizes those annotations, in a pool of multiple sparsely annotated instances with multiple invalid pixels.

If we specifically zoom in on the training scores with the mixed annotated dataset versus the scores on training with the fully annotated set, full dense annotation appears to perform well in the presence of classes that have a clearly visible border with surrounding tissues (as for example tumor or nerve), but predicted maps tend to include multiple classes when segmenting tissues that are more intertwined with neighboring tissue, as can be seen in Figure 3. Objects with a clear boundary (e.g., healthy glands) can be segmented well by most of the approaches.

Based on the proposed balancing methods on the segmentation problem at hand, we can conclude that using sparsely annotated images mixed with a little amount of densely annotated samples allows to get an overall performance that is comparable with using fully annotated instances, in particular when an instance based balancing strategy is applied. More research on different datasets, also in a field different from histopathology, is needed to verify the general validity of the proposed balancing strategies.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292, from the Dutch Cancer Society, project number 10602/2016-2, and from the Alpe dHuZes / Dutch Cancer Society Fund, grant number KUN 2014-7032.

References

Oscar GF Geessink, Alexi Baidoshvili, Joost M Klaase, Babak Ehteshami Bejnordi, Geert JS Litjens, Gabi W van Pelt, Wilma E Mesker, Iris D Nagtegaal, Francesco Ciompi, and Jeroen AWM van der Laak. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cellular Oncology*, pages 1–11, 2019.

Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 262–270. Springer, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Size-constraint loss for weakly supervised cnn segmentation. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018.

Wilma E Mesker, Jan Junggeburt, Karoly Szuhai, Pieter de Heer, Hans Morreau, Hans J Tanke, and Rob AEM Tollenaar. The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage. *Analytical Cellular Pathology*, 29 (5):387–398, 2007.

Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2017.

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.

Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3): 591–604, 2014.