Annotation-cost Minimization for Medical Image Segmentation using Suggestive Mixed Supervision Fully Convolutional Networks

Yash S. Bhalgat
Department of Computer Science
University of Michigan, Ann Arbor
yashsb@umich.edu

Meet P. Shah
Department of Electrical Engineering
Indian Institute of Technology - Bombay
meetshah1995@ee.iitb.ac.in

Abstract

For medical image segmentation, most fully convolutional networks (FCNs) need strong supervision through a large sample of high-quality dense segmentations, which is taxing in terms of costs, time and logistics involved. This burden of annotation can be alleviated by exploiting weak inexpensive annotations such as bounding boxes and anatomical landmarks. However, it is very difficult to *a priori* estimate the optimal balance between the number of annotations needed for each supervision type that leads to maximum performance with the least annotation cost. To optimize this cost-performance trade off, we present a budget-based cost-minimization framework in a mixed-supervision setting via dense segmentations, bounding boxes, and landmarks. We propose a linear programming (LP) formulation combined with uncertainty and similarity based ranking strategy to judiciously select samples to be annotated next for optimal performance. In the results section, we show that our proposed method achieves comparable performance to state-of-the-art approaches with significantly reduced cost of annotations.

1   Introduction

Image segmentation is ubiquitous in many medical image analysis applications. Unlike natural image analysis, obtaining high-quality per pixel annotations for medical images is relatively cumbersome due to the high costs, time and logistics involved. Previous works attempt to alleviate this annotation burden by either judiciously sampling the labelled datapoints in an active learning setting [2] or by leveraging weak, inexpensive forms of annotations such as bounding boxes, anatomical landmarks to assist segmentation in a mixed-supervision setting [4].

In active learning settings, a FCN starts with an initial set of images with corresponding annotations and iteratively suggests images in the remaining unlabelled dataset to be next annotated. Annotations are requested on images for which the network is highly uncertain about it's predictions and which are dissimilar to the already annotated images until a predefined annotation budget is reached.

In mixed-supervision settings, a FCN performing dense segmentation is augmented with auxiliary tasks like object detection and anatomical landmark localization which need inexpensive annotations to assist the base task of segmentation. Multi-task networks in [3, 4] have shown that adding complementary tasks enables the network learn better features and hence perform better on base task.

Costs involved in annotating each form of supervision are significantly different and therefore knowing the optimal balance between the number of annotations needed for each supervision type is essential for cost-effective segmentation. For practical applications, it would be good to have a budget-constraint framework that enables suggestive annotation in mixed supervision settings.

We propose a linear programming (LP) inspired cost-minimization framework to enable suggestive budget-constraint segmentation in mixed supervision settings.

2   Methodology

We have 3 modes of supervision in this particular application namely - dense segmentation (s), landmarks (l), and bounding box detections (d). We use the MS-Net [4] architecture which is constructed using two types of components: (i) a base network for full-resolution feature extraction and (ii) 3 sub-network extensions - one for each of $\{s, l, d\}$. We modify each sub-network to include a concrete dropout layer [1] to enable uncertainty estimation.

In the proposed method, every mode of supervision has a cost of annotation $c \in \{C_d, C_s, C_l\}$ and a value $v \in \{V_d, V_s, V_l\}$ which is an estimate of marginal gain in IoU by adding one sample of an annotation type. The training starts with a subset of the training data which has all three annotations. Let this set be denoted by $X_{init} = \{(x_1, y_{1s}, y_{1l}, y_{1d}), (x_2, y_{2s}, y_{2l}, y_{2d}), \dots\}$, where y_{is}, y_{il}, y_{id} denote the 3 annotations for the ith image. We train the network on X_{init} using all 3 of its sub-network extensions and progressively suggest images in the remaining unlabelled set to be next annotated. We call this approach Suggestive Mixed Supervision Network (SMS-Net).

Uncertainty and similarity estimation   After initial training, we obtain predictions on the remaining images in the training set. Using the concrete dropout layer during inference, for each image x from the remaining set, we compute an estimate of uncertainty for the predictions across each supervision mode - $u_s(x), u_l(x)$ and $u_d(x)$. Let $h(.)$ denote the features from the 2nd last layer of the full-resolution stream of the MS-Net. We also compute the similarity of x with X_{init} as $S(x, X_{init}) = \max_{y \in X_{init}} \kappa(h(x), h(y))$, where κ is a gaussian kernel similarity measure.

LP Formulation   The objective is to maximize the value of annotation given a cost budget. We also want to select the samples with high prediction uncertainty and the ones which are diverse from the current training set. We define a vector $z(x)$, where $z_i(x) = 1$ if we provide the ith type of annotation to image x. To select the images from the remaining set and the annotation types to be requested for each image, we solve the following modified 0-1 integer LP problem:

$$\underset{z(x), x \in X}{\arg\max} \quad \sum_{x \in X} \underbrace{V^T z(x)}_{\text{value gain}} + \underbrace{\lambda_u u(x)^T z(x)}_{\text{uncertainty}} - \underbrace{\lambda_s s(x)}_{\text{similarity}} \quad \text{subject to:} \sum_{x \in X} C^T z(x) \leq b$$

where $V = (V_s, V_d, V_l)$, $C = (C_s, C_d, C_l)$, $u = (u_s, u_d, u_l)$ and b is the allocated budget for this update. We used the `cvxpy` package for this optimization. We obtain the z vector for all the images in the remaining set. Depending on the 1's in $z(x)$, we annotate the image x, add it to X_{init} and retrain the network with this updated training set. As shown in the results section, this cost minimization formulation enables higher performance at significantly reduced annotation budgets. In each LP update, we update the estimates of V_i depending on the ΔIoU obtained. Suppose $|Y_i^{t+1}|$ is the number of samples that were given the annotation i in the $(t+1)^{th}$ LP update. Then,

$$V_i(t+1) = \alpha_i V_i(t) + (1 - \alpha_i)\beta_i \Delta IoU \quad i \in \{s, l, d\}$$

$$\alpha_i = \frac{|Y_i^t|}{|Y_i^t| + |Y_i^{t+1}|}, \beta_i = \frac{V_i |Y_i^{t+1}|}{V_s |Y_s^{t+1}| + V_l |Y_l^{t+1}| + V_d |Y_d^{t+1}|}$$

3   Results

In our experiments, we use the JSRT database [5] which has 247 high-resolution (2048×2048) chest radiographs, each with expert segmentations, 166 landmark annotations and bounding boxes [6] covering 5 anatomical structures. We compare our proposed SMS-Net with MS-Net [4] and suggestive FCN [7]. We define the max budget as the cost required to provide all the images with all 3 types of annotations. We start by randomly selecting 20% of the training samples as X_{init}. We apply the LP update as described in section 2 at K regular intervals in the entire training process. In each LP update, we assign $(\frac{1}{K})^{th}$ of the total available budget. We set $K = 10$ in all our experiments.

In Figure 1(a), we present the mean IoU obtained for varying levels of available annotation budget. Since all annotations have different costs and values, Figure 1(b) shows the distribution of *overall*
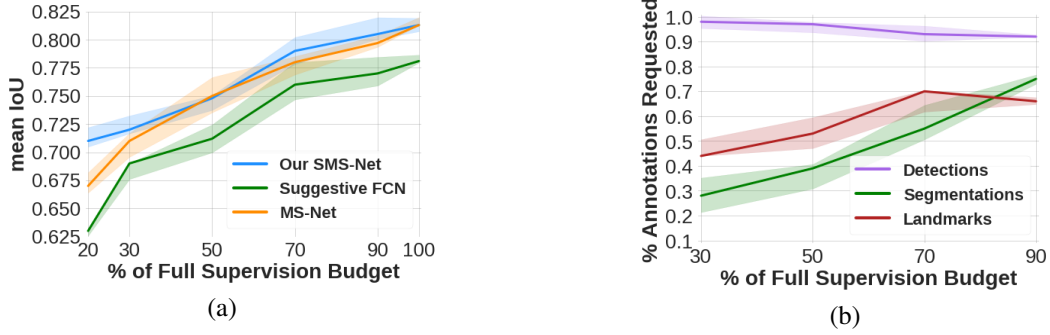
Figure 1: (a) Segmentation performance of different methods for varying levels of available budget (b) Percentage overall annotations requested by SMS-Net for varying levels of available budget. Full supervision budget is defined as the cost required to provide all the images with all 3 types of annotations

samples drawn from each annotation type for varying levels of available budget. For all experiments, we set $C = (10, 1, 7)$. These costs are chosen to be proportionally in line with practical medical image annotation scenarios. In Figure 1(b), we see that as we relax the budget constraint, more samples are provided with dense segmentations whereas lesser samples are given detections. At moderate and tighter budgets, detections and landmarks are utilized as they provide reasonable supervision at reduced costs. This selection strategy also leads to increase in performance as can be seen in Figure 1(a).

4   Conclusion and Future work

We propose a method to enable suggestive annotation in mixed supervision settings for annotation cost-minimization in medical image segmentation. We show that our method achieves better performance compared to the state-of-the-art at significantly reduced annotation budgets. Future research directions would be to evaluate the proposed method on larger datasets with a non-fixed set of annotations and to build an end to end framework for joint optimization of the base segmentation architecture and cost minimization LP.

References

[1] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017.

[2] Yingjie Gu, Zhong Jin, and Steve C Chiu. Active learning combining uncertainty and diversity for multi-class image classification. *IET Computer Vision*, 9(3):400–407, 2014.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[4] Meet P Shah, SN Merchant, and Suyash P Awate. MS-Net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 379–387. Springer, 2018.

[5] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, and Tsuneo Matsumoto. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[6] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods. *MedIA*, 10(1):19–40, 2006.

[7] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–407. Springer, 2017.