ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CVL Computer Vision Lab

Semi-supervised learning using Total variation for biomedical image segmentation

Master Thesis

Prateek Purwar
Department of Electrical Engineering and Information Technology

Advisor: Dr. Gregory Paul
Supervisor: Prof. Dr. Orcun Goksel

Abstract

Acknowledgements

I would like to acknowledge the support of my Semester Thesis advisor Dr. Gregory Paul, Post-Doc. at Computer Vision Laboratory at ETH. He consistently steered me in the right direction whenever he thought I needed it. I would sincerely thank Denis Samuylov, who helped with implementation and coding at each step of my thesis.

Contents

List of Figures

Chapter 1

Introduction

The task of image segmentation into binary classes is very useful in different cases in biomedical tasks. It can be used for detection of diseases, shape analysis etc. The methods to solve the segmentation problem has evolved among two lines: 1) level of interaction: from semi-interactive to fully automatic, and 2) level of classification: pixels to complete images. Nowadays, with the use of fully-convolutional networks, the segmentation can be obtained for complete image in one forward pass. This helps in using the local as well as contextual information for segmentation. Currently, the benchmark performance in terms of accuracy is achieved by the use of convolutional neural networks (CNN). The neural networks are specialized to learn feature maps from the examples provided and specific to task at hand. These networks require huge amount of training data: images and ground truth i.e. label for each pixel in the input image; to train the network from scratch. In literature, we can find different architectures of neural networks specially designed for task of segmentation, one of the popular architecture is U-Net[cite]. This approach works well for tasks where we can find significant number of images and can train a neural network. However, this poses a difficulty when we are trying to segment objects in microscopic images.

1.1 Electron Microscopy Images

The dataset which we are trying to segment are electron microscopic (EM) images of liver tissue. The dataset consists of a 3D stack of 2D slices of liver tissue as shown in Figure 1. The dataset can be considered as a single 3D stack or multiple 2D slices for our task. We can observe following traits in EM images:

- High variability between images: The images to be segmented may be entirely different i.e. having fixed objects as in liver tissue or having layers to segments as in neuron tissue. The objects to be segmented may differ completely from being smooth (round vesicles) to branched (neurons). This prohibits use of one dataset to train a network for another dataset and thus, restricting availability of images.
- High variability between objects to segment: The objects to be segmented vary significanly in shape, size and texture in different images.
- High variability between goals: Even for a single image, the goal of the segmentation can be totally different. The images annotated for one object can not be used again for training purpose.

These characteristics of EM images make it very difficult to fully annotate each object of interest and is extremely time consuming for experts. The difficulty to annotate different shapes and sizes can be observed

in Figure 2(a). Other difficulty for EM images is uncertainity of annotation by different experts. The difference in annotations can be observed for different experts and also for anotations of same image by same expert, as shown in Figure 2(b).

1.2 Focus of this thesis

Nowadays, it is common to train deep neural networks (DNN) using transfer learning to compensate for lack of enough data for training. Recently, Shelmar et al[cite] designed a "fully convolutional" network that take input of arbitrary size and produce segmented output for complete image. They adapted contemporary classification networks (AlexNet, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. Similar to this, we use and finetune network explained in Caelles et al[cite]. This paper tackles the task of **semi-supervised** video object segmentation, i.e., segment an object in a video, given fully annotated mask of object in the first frame. This task can be considered to be similar to segmenting objectsin 3D stack of slices. We try to finetune the network using fully annotated objects in few slices in the stack. We describe the details and observations in Section 2.

The use of pre-trained networks makes it possible to use DNN even with small amount of training data. But still to train the DNN, we need to provide fully annotated masks for objects of interest for all training images and this comes out to be a tedious and difficult task as explained above. In addition, the presence of multiple objects of different shape and sizes makes it even more diffcult and time consuming. Imagine 1000 cells in a 2D slice and possiblity to manually annotate all these cells of undefined shapes! This provide us with option of annotate few objects and train networks using either cropped images or treating rest of image as background. Or we can use semi-supervised learning using partial annotations. In literature, we can find various methods to use these partial annotation to classify each pixel as foreground or background. For example, Santner[cite] describes use of Random forests (RF) for image segmentation using partial annotations. In this thesis, we try to discover the effect of annotation budget i.e. the number of pixels to annotate and the accuracy achieved. We also try to learn which pixels to annotate to use our annotation budget efficiently.

These methods only learn pixel level information and are uncertain for maximum of pixels i.e. the probability of foreground learnt is not binary but lies between 0 and 1. In literature, different approaches can be found to use prior information to compensate for data and for the uncertainity of estimators. The most common is to use Conditional random fields (CRFs) or graph cuts to regularize the probability learnt. We solve this problem using a prior in **Bayesian framework**. Santner[cite] uses weighted total variation as prior and Random forests to learn likelihood. Ranftl[cite] uses CNN to learn unary and edge potential and combine this information to get segmentation mask using graph cuts. For our task, we implement the method described in master thesis of Dominic[cite]. In Dominic[cite], they try to learn likelihood using Random forests and prior as isotropic total variation (TV). They use a non-linear cost function to fomulate likelihood from probabilities learnt from Random forests. This is quite different to common approach of using probabilities directly as likelihood to combine with prior. Majority of researchers using CNN use a linear cost function to implement prior with help of CRFs. In this thesis, we analyze and compare these different cost functions. We try to observe the advantage of using these cost functions in different scenarios. For images as 3D stacks, it is obseved to be a difficult task and computationally efficient to encode 3D information in models as CNN or RF for learning likelihood. Also, it is common practice to use prior information in 2D. Thus, we also try to observe benefits of using 3D isotropic total variation in case of 3D stacks.

In summary, we use a Bayesian approach with RF to parametrize likelihood and isotropic TV as prior to predict segmentation mask for a given image. This gives us chance to generate fully annoatated segmentation masks and train CNN to obtain better accuracy. The common problem for use of prior is choice of appropriate scaling to couple likelihood and prior costs. Ranftl[cite] coupled the prior cost function with the likelihood cost function obtained from CNN. They optimized the final loss function to obtain optimal values for network parameters (weigts and biases) and regularization parameter. Riegler[cite] propsed a method to implement TV as specialized layers in CNN and trained the complete model, CNN + TV, together. This motivated us to replace RF with CNN and try to learn pre-trained fully convolutional network from partial annotations. We were able to restructure cross-entropy loss to compute loss for partial annotations.
Finally, we also showed advantage of using iterative semi-interactivity for efficient use of annotation budget and also to be able to provide opportunity to experts to improve learning method according to their specific requirements.

1.3 Thesis Organization

Chapter 2

Fully annotated segmentation masks

2.1 Use of pre-trained models

2.2 One shot Video object segmentation

Chapter 3

Semi-supervised image segmentation

The philosophy behind semi-supervised learning is to propagate label information from labelled to unlabelled data. Image segmentation can be seen as a classification problem which consists of assigning a class label to each pixel. For our task of binary segmentation, this means classifying each pixel as foreground or background. For our task of image segementaion, we make use of partial annotations as *scribbles*. Scribbles are pixels in image annotated by experts as foreground or background. We use example-based methods to learn from these scribbles annotated by experts. In contrast to having different images for training and testing, we use same image for training and testing as the samples used for training are pixels and not images.

3.1 Random Forest

In this section, we make use of RF as semi-supervised learning algorithm. The advantages and details of using RF can be found in Dominic[cite]. For training RF, we compute set of features in Python. We compute different features ranging from simple Sobel edge detectors to higher level Gabor filters.The choice of features was made according to WEKA[cite] toolset of FIJI[cite] plugin. These are set of 2D features and perform well medical images. We compute different type of features for a range of sigmas, which gives 69 feature maps for a single image. The details of features computed can be found in Appendix[cite]. In thesis by Dominic, we can find details and effect of feature selection for training Random forests. Here, we focus on how to get best results for given annotation budget and thus, use 20 best features computed and sufficient number of trees (30) to get best results. As shown in figure 4, we can observe that the segmentation measure saturates for more than 50 trees. We try to answer the question of where to scribble and how to make best use of our annotation budget and time.

3.1.1 Where to scribble?

In general, we believe that the more training data we provide, more we can improve the results. Does this hold for partial annotation such as scribbles? If we go on increasing the pixels annotated arbitatrily, will it improve the segmentation mask or we have to use our labelling effort intelligently to improve results? We conducted an experiment by dividing our set of foreground and background scribbles into 2 classes: easy and hard. We classified scribbles as ”easy” and ”hard” depending on effort required to annotate these pixels. For example, pixels are difficult to annotate near boundary of foreground and background, and we classify these pixels as ”hard”, as shown in figure 4. We manually scribbled image for both ”easy” and ”hard” subclasses. Then, we trained and tested RF on one image by increasing percentage of scribbles belonging

to "easy" foreground and background class. After, we have used all scribbles belonging to "easy"class, we added scribbles from "hard" class for both foreground and background. The increment was done w.r.t. total amount of scribbles, we are having and also, for higher percentage of added scribbles, we maintained a ratio between foregroun and background. The result can be observed in figure 5.

We can observe that after [cite] pixels selected from "easy" foreground and background, the segmentation measure do not improve significantly. An improvement can be observed, once we started adding "hard" scribbles. This shows that the best results can be obtained by adding "hard" scribbles after addition of certain percentage of "easy" scribbles. Looking at the plot, one might think to start adding the "hard" scribbles after [cite] "easy" scribbles. We tried this and results can be observed in figure 6.

In figure 6, as we started adding "hard" scribbles after [cite] percentage of "easy" scribbles, instead of observing a rise with additional scribbles., we observed a fall in perfomnace. This may be due to lack of enough "easy" scribbles and RF starts training its trees to focus more on "hard" scribbles. Thus, the question arises how to decide the point of addition of "hard" scribbles.

3.1.2 Iterative semi-interactive approach

In previous section, we showed need of using our annotation budget intelligently to get best performance. But, we observed the problem of deciding on how many "easy" and "hard" scribbles are needed to achieve best results. For our problem, we divided the scribbles as "easy" and "hard" according to labelling effort, but this division for scribbles may not be same from point of view of Random forest. Apriori, we don't know which pixels will be difficult for Random forest to classify correctly. The above mentioned two problems can be solved by annotating pixels iteratively to improve results, atleast once to understand which pixels are difficult for RF to classify. We show the improvement in result by doing one iteration in figure 7. We can observe the improvement in results from [cite] to [cite] for increasing the annotation budget from [cite] to [cite].

3.1.3 Uncertainity of classifier

The use of iterative semi-interactivity gives best result for given annotation budget, but, the output of RF is noisy and uncertain. The uncertainity lies in inability to classify maximum of pixel as foreground and background, as shown in figure 8(a). In figure 8(b), we can observe different results for different threshold applied on output from RF. RF acts as an classifier and classifies each pixel but we need to group these pixels into objects for segmentation. In this thesis, we make use of prior information to compensate for lack of enough annotated data and for uncertainity of classifier.

3.2 Bayesian Formulation

To make use of prior, we model our image segmentation problem as a Bayesian inference problem. Let us cosider an observed image, $\boldsymbol{I}$ and labeled or segmented groud truth, $\boldsymbol{M}$, the joint probabilty can be defined as:

$$\boldsymbol{p}(\boldsymbol{I}, \boldsymbol{M}) = \boldsymbol{p}(\boldsymbol{M})\boldsymbol{p}(\boldsymbol{I}|\boldsymbol{M}) \quad ,$$

and applying Bayes theorem,

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{M}|\boldsymbol{I}) &= \frac{\boldsymbol{p}(\boldsymbol{M})\boldsymbol{p}(\boldsymbol{I}|\boldsymbol{M})}{\boldsymbol{p}(\boldsymbol{I})} \\ &\alpha\, \boldsymbol{p}(\boldsymbol{M})\boldsymbol{p}(\boldsymbol{I}|\boldsymbol{M}) \end{aligned}$$

The left hand side is the probability of obtaining segmentation mask, $\boldsymbol{M}$ given the image $\boldsymbol{I}$, is called the posterior probability. $\boldsymbol{p}(\boldsymbol{M})$ is the prior probability of mask, $\boldsymbol{M}$. The Maximum a posteriori (MAP) estimate, $\boldsymbol{M^*}$ can be calculated as follow:

$$\boldsymbol{M^*} = \arg\max_{\boldsymbol{M}}(\boldsymbol{p}(\boldsymbol{M})\boldsymbol{p}(\boldsymbol{I}|\boldsymbol{M})) \quad . \tag{3.1}$$

The above problem can as well be stated as an energy minimization problem by writing Equation 3.1 in terms of energy:

$$\begin{aligned} E(\boldsymbol{M}) = & -\log(\boldsymbol{p}(\boldsymbol{I},\boldsymbol{M})) \\ = & -\log(\boldsymbol{p}(\boldsymbol{I}|\boldsymbol{M})) - \log(\boldsymbol{p}(\boldsymbol{M})) \\ = & E_d(\boldsymbol{I},\boldsymbol{M}) + E_r(\boldsymbol{M}) \end{aligned}$$

The total energy, E, that we want to minimize can be considered as linear combination of data or likelihood term, E_d and prior term (or regularization), E_r. This modifies calculating MAP estimate to:

$$\boldsymbol{M^*} = \arg\min_{\boldsymbol{M}}(E_d(\boldsymbol{I},\boldsymbol{M}) + E_r(\boldsymbol{M})) \quad .$$

To obtain MAP estimate, we need to formulate likelihood term and prior term. We formulate the prior using Total variation(TV). We can find use of different TV priors such as Wulff shapes etc. In our thesis, as the objects we need to segment are smooth and shaped like a circle, we make use of isotropic total variation, TV. Also, we try to use isotropic total variation in 2D and 3D as the data we are trying to segment is a 3D stack. For likelihood term, G. Paul et al.[cite] proposed an energy formulation which is not derived from a statistical model but learnt from training set. This gives the advantage of combining example-based and model-based approaches. Similar to Dominic[cite], we formulate the likelihood term in terms of a cost function, C, of soft mask, $\boldsymbol{P}$(probability of each pixel being foreground) learnt from RF. The energy minimization problems becomes:

$$\begin{aligned} E(\boldsymbol{M}) &= E_d(\boldsymbol{I},\boldsymbol{M}) + E_r(\boldsymbol{M}) \\ &=< C(\boldsymbol{P}), \boldsymbol{M} > +\lambda\, TV(\boldsymbol{M}) + \mathrm{i}_{[0,1]}(\boldsymbol{M}) \quad , \end{aligned}$$

where $\mathrm{i}_{[0,1]}(\boldsymbol{M})$ is an indicator function to ensure values of $\boldsymbol{M}$ remain in [0,1]. In addition to use of cost function, we enforce constraint of preserving the pixels annotated by experts as foreground and background in our energy minimization problem. We used an indicator function, $\mathrm{i}_{fg}(\boldsymbol{M})$, to ensure pixels in foreground scribbles have value of 1 in mask, and indicator function, $\mathrm{i}_{bg}(\boldsymbol{M})$, to ensure pixels in background scribbles have value of 0 in mask. Using discrete implementaion of TV, the final energy minimization problem is formulated as given below:

$$E(\boldsymbol{M}) =< C(\boldsymbol{P}), \boldsymbol{M} > +\lambda \, \||\boldsymbol{D}\boldsymbol{M}|_2\|_1 + \mathrm{i}_{[0,1]}(\boldsymbol{M}) + \mathrm{i}_{fg}(\boldsymbol{M}) + \mathrm{i}_{bg}(\boldsymbol{M})$$

The optimization problem is solved using Alternating Split Bregman method (ASB), as described in Dominic[cite]. The advantage of using ASB is that it splits the above problem into subproblems. Each subproblem is easy to solve and can be solved independently. The final solution to the problem is obtained by iterating updates. The details of implementation and solution can be obtained from Dominic[cite]. The results for RF with variational segmentation can be seen in figure 9. In the following section, we compare use of different cost functions in Variational segmentation methods.

3.3 Different likelihhood formulations

In literature, people have formulated the cost function either directly using the soft mask ($\boldsymbol{P}$) obtained from RF or, some linear or non-linear function of the mask. Santner[cite] formulates likelihood term as linear function of mask, C_l, as given below:

$$C_l(\boldsymbol{P}) = -4\,(\boldsymbol{P} - 0.5) \quad .$$

, while Dominic[cite] formulates it as an nonlinear (anti-nll) function. soft mask. In the following section, we compare these different formulations. For the rest of thesis, we make use of formulation described in Domimic[cite].

3.4 Effect of regularisation parameter

3.5 2D vs 3D Total Variation

Chapter 4

Semi-interactive segmentation

Chapter 5

Training CNN from scribbles

Chapter 6

Conclusion

Bibliography