A Deep Learning Approach for Semantic Segmentation in Histology Tissue Images

Jiazhuo Wang[1(✉)], John D. MacKenzie[2], Rageshree Ramachandran[3], and Danny Z. Chen[1]

[1] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, USA
jwang12@nd.edu
[2] Department of Radiology and Biomedical Imaging, UCSF, San Francisco, USA
[3] Department of Pathology, UCSF, San Francisco, USA

Abstract. To make reliable diagnosis, pathologists often need to identify certain special regions in medical images. In inflammatory bowel disease (IBD) diagnosis via histology tissue image examination, muscle regions are known to have no immune cell infiltration, and thus are ignored by pathologists. Also, messy regions (e.g., due to distortion and poor staining) are low in diagnostic yield. Hence, excluding muscle and messy regions to focus on vital regions is crucial for accurate diagnosis of IBD. In this paper, we propose a novel deep neural network based on fully convolutional networks (FCN) to identify muscle and messy regions, in an end-to-end fashion. First, we address the challenge of having limited medical training data, for training our deep neural network (a common problem for medical image processing, which may impede the application of the powerful deep learning method). Second, to deal with target regions of largely different sizes and arbitrary shapes, our deep neural network explores multi-scale information and structural information. Experimental results on clinical images show that our approach outperforms the state-of-the-art FCN for semantic segmentation of muscle and messy regions. Our approach may be readily extended to identify other types of regions in a variety of medical imaging applications.

1 Introduction

When pathologists analyze diseases by examining medical images, they often need to first identify certain special regions. In histology tissue images for inflammatory bowel disease (IBD), muscle regions are ignored by pathologists when they search for the presence of an inflammatory process, because immune cells are unlikely to infiltrate in muscle regions. Also, the manual tissue slide preparation process may create messy regions in which biological structures (e.g., various

This research was supported in part by NSF Grant CCF-1217906 and CCF-1617735, a grant of the National Academies Keck Futures Initiative (NAKFI), and NIH grant K08-AR061412-02 Molecular Imaging for Detection and Treatment Monitoring of Arthritis.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 176–184, 2016.
DOI: 10.1007/978-3-319-46723-8_21

types of immune cells) are usually not discernible or differentiable. Thus, excluding these unimportant/confusing regions may enhance computer-aided diagnosis of IBD by focusing on the critical regions (for e.g., cells identification). This poses a semantic segmentation problem (Fig. 1(a)): Assign to each pixel one of four class labels, messy region, muscle region, messy + muscle region (some muscle regions may appear to be messy; thus they are not exclusive), and background (i.e., critical regions).

In this paper, we propose a deep learning approach based on the state-of-the-art fully convolutional networks (FCN) [9] to solve this semantic segmentation problem on histology tissue images in an end-to-end fashion. In order to do so, we must overcome two major technical roadblocks.

(a) (b)

Fig. 1. (a) Left: An H&E stained histology tissue image; Right: ground truth for messy regions (dark blue), muscle regions (light blue), messy + muscle regions (yellow), and background (cyan). Note that finding the empty spaces does not require a sophisticated method. (b) Some natural scene images from Pascal VOC 2012 [6].

The first roadblock is that training deep neural networks (including FCN) usually requires a very large amount of data, in order to avoid/alleviate over-fitting. However, it is quite common to have only limited training data in medical imaging settings. U-Net [10] applied deformations to available training images to generate more data. But, it is unclear what types of deformations are best suited for each specific medical imaging modality. Further, training a model still takes lots of time. In the general computer vision (CV) community, transfer learning [4] is often applied to alleviate over-fitting and speed up training. But, medical images (including our histology tissue images) seem to be substantially different from the natural scene images used in general CV datasets (see Fig. 1(b)). Would transfer learning still be helpful to medical image processing problems (including ours)? Note that, if doing so, the source domain (working on natural scene images) and target domain (working on medical images) would be very different. Interestingly, we are able to provide an affirmative answer to this question!

The second roadblock is that our target regions can have largely varied sizes and arbitrary shapes, which we handle with two ideas. (1) We incorporate multi-scale information into our deep neural network. FCN [9] and DAG-CNN [12] used

skips to propagate low/middle level information in early layers to later layers that contain only high level information. But, a key limitation of such within-network incorporation of multi-scale information is that the scales are constrained by the size of the receptive field (RF) of the network. Therefore, we propose to utilize separate networks each incorporating information of a specific scale. In [5], a same input image was fed to separate networks with different RF sizes. We achieve the same effect (but more efficiently) by first resizing the same image into different scales, and then feeding the resized images to separate networks with identical architecture (thus of the same RF size). Note that it might be tempting to share weights among the corresponding layers of such separate networks, as in [2]. Interestingly, we show that it is more beneficial by **not sharing** such weights. (2) We incorporate structural information using conditional random field (CRF). While it is known that CRF improves the performance of FCN [1,13] which is essentially single-scale, we explore whether CRF can also boost the performance when multi-scale information is incorporated, and show that the outcome actually depends on whether weight sharing is utilized.

Experimental results on clinical data show that our approach outperforms FCN for semantic segmentation of muscle and messy regions. The results also validate our main ideas: (1) Transfer learning can help training in medical settings, even when the source and target domains are very different; (2) incorporating multi-scale information in a judicious manner boosts the performance of FCN.

2 Methodology

This section presents our multi-scale network based on FCN [9], training of the multi-scale FCN, especially wrt. transfer learning and weight sharing, and influence of CRF under the framework of multi-scale FCN.

We briefly review FCN. FCN improves over convolutional neural networks (CNN) on attaining pixel-level classification. FCN converts the fully connected layers of CNN to convolutional layers, to reduce the redundant computation incurred by overlapping sliding windows. But, the size of the score map is still smaller than input image. FCN further concatenates deconvolutional layers to up-sample score map to the size of the input image. The FCN so far is known as FCN-32s (its score map is at 32 stride), which cannot delineate object boundaries very well, because it contains only coarse information. FCN-16s improves by propagating finer scale information contained in the pool4 layer to later layers. FCN-8s propagates even finer scale information in the pool3 layer, in addition to the pool4 layer. Namely, FCN-16s and FCN-8s can be viewed as containing certain, but limited (as illustrated below), multi-scale information.

2.1 Multi-scale Information

Motivations. One target class may be more easily identified on a certain scale than other scales, and the best scales for different target classes may vary. Or, no best scale is possible, thus one has to fuse information from various scales to make

decisions. Although FCN-16s and FCN-8s already use within-network multi-scale information, the size of the receptive field (RF) of the network actually imposes some constraint. The fc7 layer of FCN (specifically, VGG-16) sees the widest range in the image, and its RF is 404 × 404 pixels theoretically (the empirical scale is actually much smaller [8]). That is, there is no way for FCN-16s or FCN-8s to see a wider range than 404 × 404 pixels, even if it might be beneficial by doing so. Hence, we propose to incorporate multi-scale information, by using separate networks each covering a specific field of view (FOV) in the image.

Architecture of Multi-scale FCN. Our main idea (Fig. 2) is to first apply various FCNs, each of which takes care of a different FOV in the input image. Then, we fuse the score maps (SMs) produced by those FCNs. Finally, our fused score maps will go through a soft-max function [7] to compute a cross entropy classification loss [7]. Below, we discuss several key aspects in more detail.

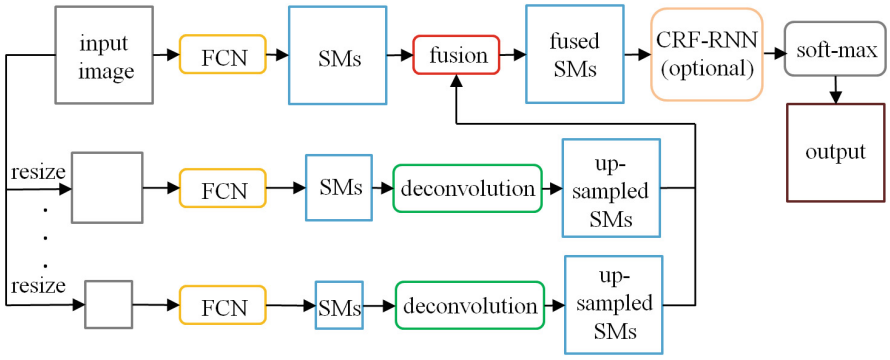

Fig. 2. The architecture of our multi-scale FCN.

To make the fc7 layer of FCN see wider, one design choice [5] is to change the hyper-parameters (e.g., the filters size), which increases too much computational burden. Instead, we shrink input image (by bilinear interpolation) to smaller sizes, and feed resized images to various FCNs with an identical structure. In this way, although the fc7 layer of each FCN still has a fixed receptive field of 404 × 404 pixels in the shrunk images, it actually sees a wider range in the original input image (before resizing). Using an identical structure for various FCNs also makes apply transfer learning easily, using the pre-trained FCN [9].

Note that shrinking the input image to a smaller size version will make the score maps of each FCN be of the same size as that resized image. Hence, we need to add another deconvolutional layer after the original FCN structure to up-sample the score maps thus obtained to the size of the original input image.

During the fusion step, we simply sum up the values at the corresponding locations of the score maps from various FCNs. To make an end-to-end system, we use a during-training fusion, as opposed to a post-training fusion. Specifically,

during-training fusion allows the update of the parameters of each FCN to be influenced by other FCNs. Namely, parameters of various FCNs are learned in a correlated way, due to such mutual influence. One advantage of this is that one FCN having a wider view could be viewed as a context that regulates another FCN having a narrower view; an FCN having a narrower view improves the ability of delineating finer boundaries for another FCN having a wider view. Post-training fusion means that each FCN learns separately during training (by computing a separate cross entropy loss for each FCN). That is, there is no communication or mutual influence between FCNs having different view ranges.

2.2 Training

We apply a stochastic gradient decent (SGD) algorithm [7] to learn the parameters in our network. We explore two key aspects for training, parameter initialization (via transfer learning), and parameter update (via weight sharing).

Transfer Learning. It is common only limited training data is available in medical image processing. But, deep networks normally require a very large amount of training data, and the training process usually takes a long time even on modern GPUs. Our main idea for this is to apply transfer learning (TL) [4].

The essential idea of TL is that learning a new task can be facilitated by transferring relevant knowledge from a related task that has already been learned. Two networks are involved in TL, the source network (S-net) and target network (T-net). The T-net is for the new task, trained on datasets that one currently has (in our case, histology tissue images). The S-net has already been trained for a related task on some other datasets (in our case, we use the pre-trained FCN in [9]). The knowledge is transferred from S-net to T-net by initializing the parameter values in T-net as the corresponding parameter values in S-net.

It is natural to think that there should be some domain similarity between the new task and the related task, in order to make TL work well. Namely, images for the related task should look similar enough to those for the new task. But, histology tissue images are drastically different from natural scene images. An immediate concern is whether TL still helps. The answer turns out to be "yes". Our intuition is: The difference between these two image domains is mainly at the high semantic level; nevertheless, the two domains still share some common properties at the low, middle level image cues (such common properties, like edges, corners, and correlation between them, may not be easily observed by human eyes, whereas high level features are more attractive to human eyes).

Weight Sharing. Weight sharing (WS) is commonly applied when one uses multiple networks with identical structures. Namely, the corresponding parameters (i.e., weights) in such networks share common values during the learning process. For example, the Siamese network [3] applied WS to its two CNNs to learn a similarity metric for a pair of input images. Recurrent neural networks (RNN) [11] can be viewed as applying WS to networks for different time steps.

It might be tempting to apply WS to the multiple FCNs in our network, as in [2]. However, doing so would make the learned shared parameters capture only

scale-independent information, and lose some scale-specific information. On the contrary, with no WS applied, each FCN would be specialized on a certain scale, and together, the FCNs would collect all information from multiple scales. Our experimental results empirically show semantic segmentation benefits more from multi-scale information than from merely scale-independent information.

2.3 Structural Information

Conditional random field (CRF) was applied in [1] as a post-processing step after FCN to incorporate structural information. Namely, CRF uses the probabilities produced by FCN as its unary cost, and it also considers pairwise cost imposing smoothness and consistency for label assignments. In [13], CRF was implemented as an RNN (called CRF-RNN), so that CRF can be jointly trained with FCN.

We examine the influence of structural information in the context of multi-scale FCN. We place CRF-RNN after the fusion step from various FCNs (i.e., the fused score maps will be used as unary energy for CRF), and before the soft-max function (see Fig. 2). Given that we have incorporated multi-scale information (specifically, the one without weight sharing), we find CRF is not as helpful as in the case for single-scale FCN. Our intuition for this is that since the FCNs seeing wider regulate the FCNs seeing narrower, such regulation can be viewed as similar smoothness constraint provided by pairwise energy of CRF. However, if weight sharing is applied, then we find CRF can still improve the performance, which indicates that weight sharing may make such regulation effect weaker.

3 Experiments and Discussions

We collected clinical H&E stained histology tissue whole slides (originally scanned at 40X magnification, then resized to 10X to save computational costs). We cut whole slides into images of size 1000×1000 pixels, due to memory constraint of Caffe implementation [7]. We sampled from them nearly 200 images to manually mark the ground truth data at the pixel level based on histology criteria.

We use 2-fold cross validation to evaluate two standard metrics for semantic segmentation: Pixel accuracy (*pixel-acc*) and region intersection over union (*IU*), defined as follows. Let n_c denote the number of target classes and n_{ij} denote the number of pixels of class i predicted as class j. Then *pixel-acc* $= \sum_i n_{ii} / \sum_i \sum_j n_{ij}$, and $IU = (1/n_c) \sum_i (n_{ii} / (\sum_j n_{ij} + \sum_j n_{ji} - n_{ii}))$.

We use FCN-16s [9] as baseline for comparison, and as the basic component for each scale of our network (we found both FCN-8s and FCN-32s decrease the performance, compared to FCN-16s, on our data). We evaluate the key factors (transfer learning (TL), multi-scale (MS) information, weight sharing (WS), and CRF) in a controlled and structured manner as below. The quantitative results and some visual results are shown respectively in Table 1 and Fig. 3.

Transfer Learning. First, we compare training from scratch to that applying TL (i.e., initializing the parameter values of FCN-16s by the pre-trained network

Table 1. Quantitative performance of different methods.

Method	FCN	TL+FCN	TL+MS+WS	TL+MS+WS+CRF	TL+MS	TL+MS+CRF
pixel-acc	0.73	0.82	0.85	0.88	**0.90**	**0.90**
IU	0.39	0.45	0.48	0.50	**0.56**	0.54

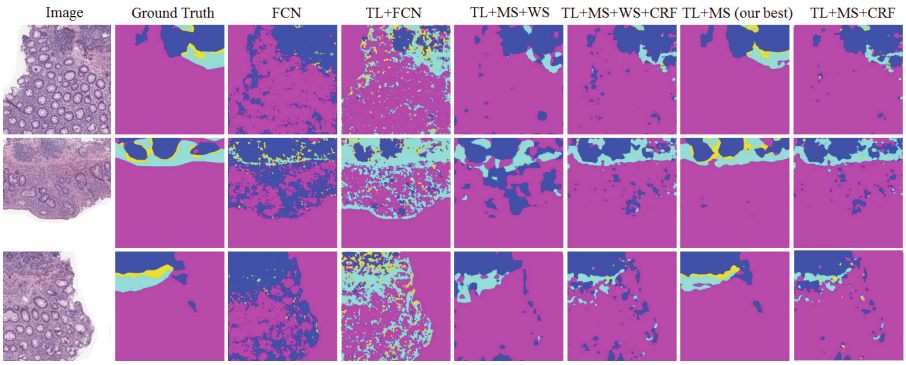

Fig. 3. Examples of visual results for different methods.

[9], and fine-tuning the model using our histology images). For training from scratch (trained nearly 40000 iterations), its learning rate and momentum are set respectively as 10^{-9} and 0.90; for TL (trained less than 2000 iterations), they are respectively 10^{-11} and 0.99 (these values are used throughout the experiments for other TL related methods). As shown in Table 1, FCN+TL improves the performance of FCN significantly. This validates that TL can still be helpful (i.e., learning a good model quickly), even if the domain difference is drastic.

Multi-scale Information. Second, we examine the influence of incorporating multi-scale information by various FCNs. The relevant results shown in Table 1 are based on only two FCNs. The first FCN takes the original 1000 × 1000 size image as input; the second takes a resized image of size 500 × 500.

Note that both TL+MS and TL+MS+WS (i.e., regardless of whether WS is applied) outperform TL+FCN. We were curious whether the improvement is due to multiple FCNs, or just the additional FCN (taking resized input). Thus, we trained a slightly different version of TL+FCN, taking an input image of size 500 × 500, instead of 1000 × 1000. We found this new version performs a little worse than previously. This suggests the improvement is due to multiple FCNs.

We also evaluated a three-FCN model, by adding a third FCN taking a resized image of size 250 × 250. This improves only slightly over the two-FCN model, probably because the third FCN contains information too abstract to be useful.

Weight Sharing. Third, we evaluate whether weight sharing should be applied. As shown in Table 1, regardless of whether CRF is used, TL+MS+WS is worse than its counterpart without weight sharing, TL+MS; also, TL+MS+WS+CRF is worse than TL+MS+CRF. This implies that it is better off to let each

individual FCN be specialized at a certain scale of information, as apposed to extracting merely scale-independent information from all the FCNs using weight sharing.

Structural Information. At last, we examine the effect of incorporating structural information, given that multi-scale information has been incorporated. Table 1 shows that TL+MS+WS+CRF outperforms its counterpart without structural information, TL+MS+WS; but, TL+MS+CRF performs similarly as TL+MS. This suggests that as long as multi-scale information is incorporated appropriately, the additional structural information may not be very useful. A possible explanation for this is that the FCNs seeing wider impose on the FCNs seeing narrower a similar consistency constraint as that from CRF.

4 Conclusions

In this paper, we propose a new deep learning approach for semantic segmentation of messy and muscle regions in histology tissue images. We show that (1) transfer learning can help training effectively, even when the differences between the source domain and target domain seem very large; (2) incorporating multi-scale information in an appropriate way can greatly improve the performance.

References

1. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2014)
2. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. CoRR (2015)
3. Chopra, S., Hadsell, R., Lecun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR, pp. 539–546 (2005)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. In: ICML, pp. 647–655 (2014)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2012, VOC 2012 Results (2012). http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
7. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrel, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint (2014). arXiv:1408.5093
8. Liu, W., Rabinovich, A., Berg, A.C.: ParseNet: looking wider to see better. CoRR (2015)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)

10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_28
11. Sutskever, I.: Training recurrent neural networks. Ph.D. thesis (2012)
12. Yang, S., Ramanan, D.: Multi-scale recognition with DAG-CNNs. In: ICCV (2015)
13. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.: Conditional random fields as recurrent neural networks. In: ICCV (2015)