Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation

Lin Yang[1], Yizhe Zhang[1], Jianxu Chen[1], Siyuan Zhang[2], Danny Z. Chen[1]

[1] Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, IN 46556, USA
[2] Department of Biological Sciences, Harper Cancer Research Institute,
University of Notre Dame, Notre Dame, IN 46556, USA

Abstract. Image segmentation is a fundamental problem in biomedical image analysis. Recent advances in deep learning have achieved promising results on many biomedical image segmentation benchmarks. However, due to large variations in biomedical images (different modalities, image settings, objects, noise, etc), to utilize deep learning on a new application, it usually needs a new set of training data. This can incur a great deal of annotation effort and cost, because only biomedical experts can annotate effectively, and often there are too many instances in images (e.g., cells) to annotate. In this paper, we aim to address the following question: With limited effort (e.g., time) for annotation, what instances should be annotated in order to attain the best performance? We present a deep active learning framework that combines fully convolutional network (FCN) and active learning to significantly reduce annotation effort by making judicious suggestions on the most effective annotation areas. We utilize uncertainty and similarity information provided by FCN and formulate a generalized version of the maximum set cover problem to determine the most representative and uncertain areas for annotation. Extensive experiments using the 2015 MICCAI Gland Challenge dataset and a lymph node ultrasound image segmentation dataset show that, using annotation suggestions by our method, state-of-the-art segmentation performance can be achieved by using only 50% of training data.

1 Introduction

Image segmentation is a fundamental task in biomedical image analysis. Recent advances in deep learning [2,3,12,15,16] have achieved promising results on many biomedical image segmentation benchmarks [1,14]. Due to its accuracy and generality, deep learning has become a main choice for image segmentation. But, despite its huge success in biomedical applications, deep learning based segmentation still faces a critical obstacle: the difficulty in acquiring sufficient training data due to high annotation efforts and costs. Comparing to applications in natural scene images, it is much harder to acquire training data in biomedical applications for two main reasons. (1) Only trained biomedical experts can annotate data, which makes crowd leveraging quite difficult. (2) Biomedical images often contain much more object instances than natural scene images, which

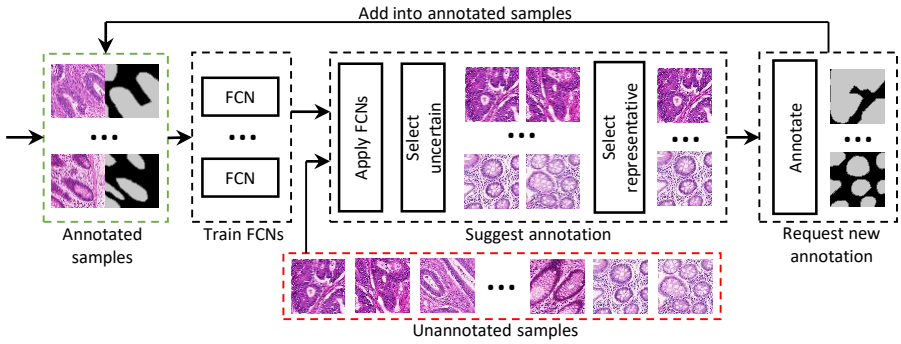

Fig. 1. Illustrating our overall deep active learning framework.

can incur extensive manual efforts of annotation. For example, public datasets in biomedical areas have significantly fewer spatial annotated images (85 for MICCAI Gland Challenge [14]; 30 for ISBI EM Challenge [1]).

To alleviate the common burden of manual annotation, an array of weakly supervised segmentation algorithms [8] has been proposed. However, they did not address well the question that which data samples should be selected for annotation for high quality performance. Active learning [13], which allows the learning model to choose training data, provided a way to answer this need. As shown in [10], using active learning, state-of-the-art level performance can be achieved using significantly less training data in natural scene image segmentation. But, this method is based on the pre-trained region proposal model and pre-trained image descriptor network, which cannot be easily acquired in biomedical image settings due to large variations in biomedical applications.

In this paper, we present a new framework that combines fully convolutional network (FCN) [11] and active learning [13] to reduce annotation effort by making judicious suggestions on the most effective annotation areas. To address the issues in [10], we exploit FCN to obtain domain specific image descriptor and directly generate segmentation without using region proposals. Fig. 1 outlines the main ideas and steps of our deep active learning framework. Starting with very little training data, we iteratively train a set of FCNs. At the end of each stage, we extract useful information (such as uncertainty estimation and similarity estimation) from these FCNs to decide what will be the next batch of images to annotate. After acquiring the new annotation data, the next stage is started using all available annotated images. Although the above process seems straightforward, we need to overcome several challenges in order to integrate FCNs into this deep active learning framework, as discussed below.

Challenges from the perspective of FCNs. (1) The FCNs need to be fast to train, so that the time interval between two annotation stages is acceptable. (2) They need to be of good generality, in order to produce reasonable results when little training data is available. To make the model fast to train, we utilize the ideas of batch normalization [9] and residual networks [6]. Then, we use

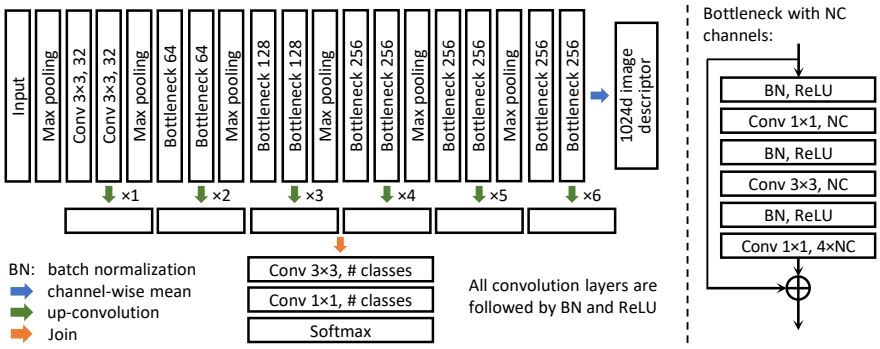

Fig. 2. Illustrating the detailed structure of our FCN components.

bottleneck design [6] to significantly reduce the number of parameters (for better generality) while maintaining a similar number of feature channels as in [3].

Challenges from the perspective of active learning. It needs to exploit well the information provided by the FCNs when determining the next batch of training data. For this, we first demonstrate how to estimate uncertainty of the FCNs based on the idea of bootstrapping and how to estimate similarity between images by using the final layer of the encoding part of the FCNs. Based on such information, we formulate a generalized version of the maximum set cover problem [5,7] for suggesting the next batch of training data.

Experiments using the 2015 MICCAI Gland Challenge dataset [14] and a lymph node ultrasound image segmentation dataset [17] show that (1) annotation suggestions by our framework are more effective than common methods such as random query and uncertainty query, and (2) our framework can achieve state-of-the-art segmentation performance by using only 50% of training data.

2 Method

Our proposed method consists of three major components: (1) a new FCN, which shows state-of-the-art performance on the two datasets used in our experiments; (2) uncertainty estimation and similarity estimation of the FCNs; (3) an annotation suggestion algorithm for selecting the most effective training data.

2.1 A new fully convolutional network

Based on recent advances of deep neural network structures such as batch normalization [9] and residual networks [6], we carefully design a new FCN that has better generality and is faster to train.

Fig. 2 shows the detailed structure of our new FCN. Its encoding part largely follows the structure of DCAN [3]. As shown in both residual networks [6] and batch normalization [9], a model with these modifications can achieve the same accuracy with significantly fewer training steps comparing to its original version.

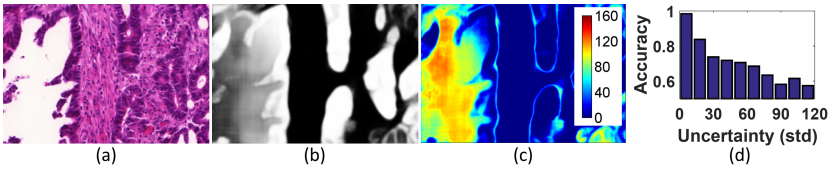

Fig. 3. (a) An original image; (b) the probability map produced by our FCNs for (a); (c) uncertainty estimation of the result; (d) relation between uncertainty estimation and pixel accuracy on the testing data. This shows that the test accuracy is highly correlated with our uncertainty estimation.

This is essential when combining FCNs and active learning, since training FCNs usually takes several hours before reaching a reasonable performance. Thus, we change the original convolution layers into residual modules with batch normalization. Note that, at the start of active learning, since only few training samples are available, having too many free parameters can make the model hard to train. Hence, we utilize the bottleneck design [6] to reduce the number of parameters while maintaining a similar number of feature channels at the end of each residual module. In the decoding part of the network, we modify the structure in [2] to gradually enlarge the size of the feature maps to ensure a smooth result. Finally, a 3×3 convolution layer and a 1×1 convolution layer are applied to combine the feature maps from different scales together. As the experiments show, our new FCNs can achieve state-of-the-art performance when all training data is used while still able to produce reasonable results when very little training data is available.

2.2 Uncertainty estimation and similarity estimation

A straightforward strategy to find the most "valuable" annotation areas is to use uncertainty sampling, with the active learner querying the most uncertain areas for annotation. However, since deep learning models tend to be uncertain for similar types of instances, simply using uncertainty sampling will result in duplicated selections of annotation areas. To avoid this issue, our method aims to select not only uncertain but also highly representative samples (samples that are similar to lots of other training samples). To achieve this goal, we need to estimate the uncertainty of the results and measure the similarity between images. In this section, we illustrate how to extract such information from FCNs.

Bootstrapping [4] is a standard way for evaluating the uncertainty of learning models. Its basic idea is to train a set of models while restricting each of them to use a subset of the training data (generated by sampling with replacement) and calculate the variance (disagreement) among these models. We follow this procedure to calculate the uncertainty of FCNs. Although the inner variance inside each FCN can lead to overestimation of the variance, in practice, it can still provide a good estimation of the uncertainty. As shown in Fig. 3(d), the estimated uncertainty for each pixel has a strong correlation with the testing errors. Thus,

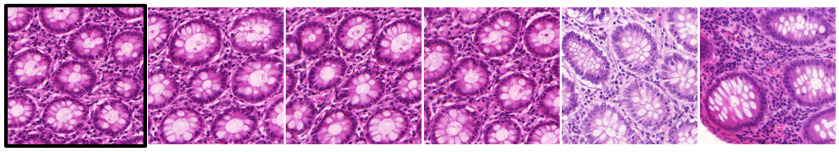

Fig. 4. Illustrating similarity estimation: The 5 images on the right have the highest similarity scores with respect to the leftmost images among all training images in [14].

selecting uncertain training samples can help FCNs to correct potential errors. Finally, the overall uncertainty of each training sample is computed as the mean uncertainty of its pixels.

CNN based image descriptor has helped produce good results in natural scene images. The encoding part of FCN is naturally an CNN, and for an input image I_i, the output of the last convolution layer in the encoding part can be viewed as high level features I_i^f of I_i. Next, to eliminate shifting and rotation variances of the image, we calculate the channel-wise mean of I_i^f to generate condensed features I_i^c as the domain-specific image descriptor. This approach has two advantages. (1) There is no need to train another separate image descriptor network. (2) Because the FCNs are trying to compute the segmentation of the objects, I_i^c contains rich and accurate shape information. Finally, we define the similarity estimation between two images I_i and I_j as: $sim(I_i, I_j) = cosine\_similarity(I_i^c, I_j^c)$. Fig. 4 shows an example of the similarity estimation.

2.3 Annotation suggestion

To maximize the effectiveness of the annotation data, the annotated areas are desired to be typical or representative in terms of the following two properties. (1) Uncertainty: The annotated areas need to be difficult to segment for the network. (2) Representativeness: The annotated areas need to bear useful characteristics or features for as many unannotated images as possible. In this section, we show how to suggest a set of areas for annotation that very well satisfy these two properties, based on similarity estimation and uncertainty estimation.

In each annotation suggestion stage, among all unannotated images, $\mathcal{S}_u$, we aim to select a subset of k images, $\mathcal{S}_a \subseteq \mathcal{S}_u$, that is both highly uncertain and representative. Since uncertainty is a more important criterion, in step 1, images with the top K ($K > k$) uncertainty scores are extracted and form a candidate set $\mathcal{S}_c$. In step 2, we find $\mathcal{S}_a \subseteq \mathcal{S}_c$ that has the largest representativeness.

To formalize the representativeness of $\mathcal{S}_a$ for $\mathcal{S}_u$, we first define the representativeness of $\mathcal{S}_a$ for an image $I_x \in \mathcal{S}_u$ as: $f(\mathcal{S}_a, I_x) = \max_{I_i \in \mathcal{S}_a} sim(I_i, I_x)$, where $sim(\cdot, \cdot)$ is the similarity estimation between I_i and I_x. Intuitively, I_x is represented by its most similar image in $\mathcal{S}_a$, measured by the similarity $sim(\cdot, \cdot)$. Then, we define the representativeness of $\mathcal{S}_a$ for $\mathcal{S}_u$ as: $F(\mathcal{S}_a, \mathcal{S}_u) = \sum_{I_j \in \mathcal{S}_u} f(\mathcal{S}_a, I_j)$, which reflects how well $\mathcal{S}_a$ represents all the images in $\mathcal{S}_u$. By finding $\mathcal{S}_a \subseteq \mathcal{S}_c$ that maximizes $F(\mathcal{S}_a, \mathcal{S}_u)$, we promote $\mathcal{S}_a$ by (1) selecting k "hub" images that

Table 1. Comparison with full training data for gland segmentation.

Method	F1 score		ObjectDice		ObjectHausdorff	
	Part A	Part B	Part A	Part B	Part A	Part B
Our method	**0.921**	**0.855**	0.904	**0.858**	44.736	**96.976**
Multichannel [16]	0.893	0.843	**0.908**	0.833	**44.129**	116.821
Multichannel [15]	0.858	0.771	0.888	0.815	54.202	129.930
CUMedVision [3]	0.912	0.716	0.897	0.781	45.418	160.347

Table 2. Results for lymph node ultrasound image segmentation.

Method	Mean IU	F1 score	Method	Mean IU	F1 score
U-Net [12]	0.798	0.775	Uncertainty 50%	0.858	0.849
CUMedNet [2]	0.816	0.798	Our method 50%	0.875	0.871
CFS-FCN [17]	0.851	0.843	Our method full	**0.879**	**0.874**

are similar to many unannotated images and (2) covering diverse cases (since adding annotation to the same case does not significantly increase $F(\mathcal{S}_a, \mathcal{S}_u)$).

Finding $\mathcal{S}_a \subseteq \mathcal{S}_c$ with k images that maximizes $F(\mathcal{S}_a, \mathcal{S}_u)$ can be formulated as a generalized version of the maximum set cover problem [5], as follows. We first show when $sim(\cdot,\cdot) \in \{0,1\}$, the problem is an instance of the maximum set cover problem. For each image $I_i \in \mathcal{S}_c$, I_i covers a subset $\mathcal{S}_{I_i} \subseteq \mathcal{S}_u$, where $I_y \in \mathcal{S}_{I_i}$ if and only if $sim(I_i, I_y) = 1$. Further, since $sim(\cdot,\cdot) \in \{0,1\}$, for any $I_x \in \mathcal{S}_u$, $f(\mathcal{S}_a, I_x)$ is either 1 (covered) or 0 (not covered) and $F(\mathcal{S}_a, \mathcal{S}_u)$ (the sum of $f(\mathcal{S}_a, I_x)$'s) is the total number of the covered images (elements) in $\mathcal{S}_u$ by $\mathcal{S}_a$. Thus, finding a k-images subset $\mathcal{S}_a \subseteq \mathcal{S}_c$ maximizing $F(\mathcal{S}_a, \mathcal{S}_u)$ becomes finding a family $\mathcal{F}$ of k subsets from $\{S_{I_i} \mid I_i \in \mathcal{S}_c\}$ such that $\cup_{S_j \in \mathcal{F}} S_j$ covers the largest number of elements (images) in $\mathcal{S}_u$ (max k-cover [5]). The maximum set cover problem is NP-hard and its best possible polynomial time approximation algorithm is a simple greedy method [5] (iteratively choosing S_i to cover the largest number of uncovered elements). Since our problem is a generalization of this problem (with $sim(\cdot,\cdot) \in [0,1]$, instead of $sim(\cdot,\cdot) \in \{0,1\}$), our problem is clearly NP-hard, and we adopt the same greedy method. Initially, $\mathcal{S}_a = \emptyset$ and $F(\mathcal{S}_a, \mathcal{S}_u) = 0$. Then, we iteratively add $I_i \in \mathcal{S}_c$ that maximizes $F(\mathcal{S}_a \cup I_i, \mathcal{S}_u)$ over $\mathcal{S}_a$, until $\mathcal{S}_a$ contains k images. Note that, due to the max operation in $f(\cdot,\cdot)$, adding an (almost) duplicated I_i does not increase $F(\mathcal{S}_a, \mathcal{S}_u)$ by much. It is easy to show that this algorithm achieves an approximation ratio of $1 - \frac{1}{e}$ [7].

3 Experiments and Results

To thoroughly evaluate our method on different scenarios, we apply it to the 2015 MICCAI Gland Challenge dataset and a lymph node ultrasound image segmentation dataset [17]. The MICCAI data have 85 training images and 80 testing images (60 in Part A; 20 in Part B). The lymph node data have 37

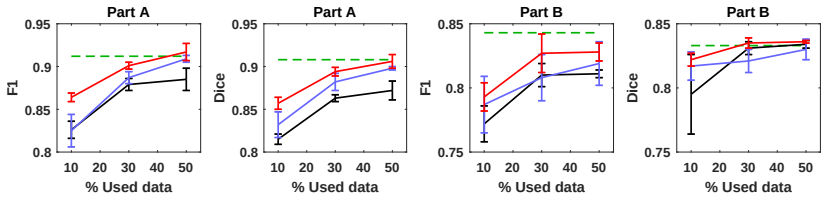

Fig. 5. Comparison using limited training data for gland segmentation: The black curves are for the results of random query, the blue curves are for the results of uncertainty query, the red curves are for the results by our annotation suggestion, and the dashed green lines are for the current state-of-the-art results using full training data.

training images and 37 testing images. In our experiments, we use $k = 8$, $K = 16$, 2000 training iterations, and 4 FCNs. The waiting time between two annotation suggestion stages is 10 minutes on a workstation with 4 NVIDIA Telsa P100 GPU. We use 5% of training data as validation set to select the best model.

Gland segmentation. We first evaluate our FCN module using full training data. As Table 1 shows, on the MICCAI dataset, our FCN module achieves considerable improvement on 4 columns ($\sim$ 2% better), while has very similar performance on the other two ($\sim$ 0.5% worse). Then, we evaluate the effectiveness of our annotation suggestion method, as follows. To simulate the annotation suggestion process, we reveal training annotation only when the framework suggests it. The annotation cost is calculated as the number of revealed pixels. Once the annotation cost reaches a given budget, we stop providing more training data. In our experiment, we set this budget as 10%, 30%, and 50% of the overall labeled pixels. We compare our method with (1) random query: randomly requesting annotation before reaching the budget, and (2) uncertainty query: selecting annotation areas based only on uncertainty estimation ($K = k$). Fig. 5 summarizes the results. It shows that our annotation suggestion method is consistently better than random query and uncertainty query, and our framework can achieve state-of-the-art performance using only 50% of the training data.

Lymph node segmentation. Table 2 summarizes the results on lymph node segmentation. "Our method full" entry shows the results of our FCN using all training data. "Our method 50%" and "Uncertainty 50%" entries show the comparison between uncertainty query and our annotation suggestion method under the 50% budget. It shows that our framework achieves better performance in all cases. By using 50% of the training data, our framework attains better segmentation performance than the state-of-the-art method [17].

4 Conclusions

In this paper, we presented a new deep active learning framework for biomedical image segmentation by combining FCNs and active learning. Our new method provides two main contributions: (1) A new FCN model that attains state-of-the-art segmentation performance; (2) an annotation suggestion approach that

can direct manual annotation efforts to the most effective annotation areas.

Acknowledgment. This research was supported in part by NSF Grants CCF-1217906, CNS-1629914, CCF-1617735, CCF-1640081, NIH Grant 5R01CA194697-03, and the Nanoelectronics Research Corporation, a wholly-owned subsidiary of the Semiconductor Research Corporation, through Extremely Energy Efficient Collective Electronics, an SRC-NRI Nanoelectronics Research Initiative under Research Task ID 2698.005.

References

1. Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Cirean, D., Giusti, A., Gambardella, L.M., et al.: Crowdsourcing the creation of image segmentation algorithms for connectomics. Frontiers in Neuroanatomy 9, 142 (2015)
2. Chen, H., Qi, X., Cheng, J.Z., Heng, P.A.: Deep contextual networks for neuronal structure segmentation. In: AAAI. pp. 1167–1173 (2016)
3. Chen, H., Qi, X., Yu, L., Heng, P.A.: Dcan: Deep contour-aware networks for accurate gland segmentation. In: CVPR. pp. 2487–2496 (2016)
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC press (1994)
5. Feige, U.: A threshold of ln n for approximating set cover. JACM 45(4), 634–652 (1998)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
7. Hochbaum, D.S.: Approximating covering and packing problems: Set cover, vertex cover, independent set, and related problems. In: Approximation Algorithms for NP-hard Problems. pp. 94–143. PWS Publishing Co. (1996)
8. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: NIPS. pp. 1495–1503 (2015)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
10. Jain, S.D., Grauman, K.: Active image segmentation propagation. In: CVPR. pp. 2864–2873 (2016)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
13. Settles, B.: Active learning literature survey. University of Wisconsin, Madison 52(55-66), 11 (2010)
14. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., et al.: Gland segmentation in colon histology images: The GlaS challenge contest. Medical Image Analysis 35, 489–502 (2017)
15. Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Chang, E.I.: Gland instance segmentation by deep multichannel side supervision. In: MICCAI. pp. 496–504 (2016)
16. Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., et al.: Gland instance segmentation using deep multichannel neural networks. arXiv preprint arXiv:1611.06661 (2016)
17. Zhang, Y., Ying, M.T., Yang, L., Ahuja, A.T., Chen, D.Z.: Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images. In: BIBM. pp. 443–448. IEEE (2016)