A Deep Fully Convolutional Network for Distal Radius and Ulna Semantic Segmentation

Shuqiang Wang1,*, Wei Liang2, Hongfei Wang1, Zhuo Chen1 and Yiqian Lu1
1Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
2Liaocheng Institute of Water Resource Survey and Design, Shandong, China
*sq.wang@siat.ac.cn

Abstract. Semantic segmentation is an essential step to do further image analysis and scene understanding tasks. In medical imaging analysis applications, it is even more challenging to do automatic segmentation due to tissues’ complicated boundaries. In this paper, a fully convolutional network (FCN) based model is constructed to segment distal radius and ulna (DRU) areas from hand X-ray images. We evaluated the proposed network on a clinical DRU dataset with different network configurations. The proposed network can achieve 98% accuracy and 96% mean Intersection over Union (IoU).

1. Introduction

The purpose of image segmentation is trying to cluster pixels into distinct regions according to interest features. Conventional segmentation methods including region-based algorithms, boundary-based algorithms and graph-based image segmentation methods. To acquire accurate segmentation results is still a challenging problem. In recent years, deep learning techniques have dramatically improved the state-of-the-art in visual object detection and recognition [1]. In medical image analysis areas, research groups have tried to apply CNNs to a broad range of applications and got promising results [2]. For example, Pereira et al. proposed an automatic brain tumors segmentation method based on convolutional neural networks. Their method achieved top results on a 2013 public challenge dataset [3]. However, CNN-based segmentation algorithms always suffer from low efficiency and low accuracy. In 2015, Long et al. proposed a full convolutional network (FCN) which can classify each pixel of an image [4]. From then on, FCNs became more popular to do semantic segmentation tasks than conventional CNN models.

To evaluate bone age in patients will guide physicians to make clinical decisions. Luk et al. have proposed a novel skeletal maturity assessment method based on distal radius and ulna (DRU) radiographs [5]. Extracting accurate DRU regions from hand X-ray images is the key step to apply this new DRU classification scheme to assess bone age. In this paper, we propose a fully convolutional network (FCN) to do distal radius and ulna semantic segmentation. The main contributions of this paper include: (1). To our best knowledge, this is the first work to apply FCN to DRU segmentation and get satisfactory results. (2). We deployed numerous experiments to investigate the proposed network performance with different network configurations and different training samples.

2. Related work

Convolutional neural networks have been widely used in medical image segmentation tasks. Karnitsas et al. proposed a 3D fully connected Conditional Random Field to post-process the network’s
soft segmentation [6]. In 2016, Moeskopset al. presented a CNN based system to automatically segment MR brain images into multiple tissue classes.[7]. Also, Roth et al. proposed a multi-level deep CNNs to do pancreas segmentation in abdominal CT scans [8]. From the above-mentioned works, it can be concluded that more CNN related methods have been employed to medical image segmentation and achieved promising results in the recent years.

In 2015, fully convolutional network (FCN) was presented to do semantic segmentation. For example, Ronnebergeret al. presented a fully convolutional network for segmentation of neuronal structures in electron microscopic stacks. Their proposed network was the winner of ISBI cell tracking challenge 2015 [9]. In 2017, Andrearczyket al. developed a fully convolutional network for texture segmentation. [10]. Moreover, FCN can be applied to video segmentation problems. Valipouret al. built a fully convolutional network and used a recurrent unit to work on temporal data [11]. These cited works have proved FCN has strong convincing performance on semantic segmentation. In this paper, we will try to apply FCN to distal radius and ulna segmentation from hand X-ray images. The proposed method and corresponding experiments will be presented in detail in the following sections.

3. Method

3.1 Fully Convolutional Neural Network and Semantic Segmentation

In convolutional neural networks, fully conventional networks (FCNs) were proposed to replace all the fully connected layers by convolutional layers to prevent spatial information damage. Both the input and output data of FCNs are two-dimensional images. The output results contain spatial structure information according to the input images. FCNs use deconvolution operations to reconstruct the output results from the final feature map which is a heatmap. Then, softmax will be applied to each pixel of generated images to do pixelwise classification. The final semantic segmentation is built on pixelwise classification results of the reconstructed image. We fuse feature maps generated from convolutional layers and feature maps from previous pooling layers together. Then the mixed feature maps will be upsampled back to the image. The purpose of these operations is to utilize the shallow and deep features for semantic segmentation.

Compared to traditional convolution neural networks, fully convolutional networks have two remarkable advantages: (1). arbitrary input size accepted: FCNs do not require training data have identical size; (2). high training efficiency: FCNs avoid repetitive computation and storage overhead of small surrounding regions. Based on the above description, the integrated working flowchart of FCNs semantic segmentation is illustrated in the Figure 1.

![Figure 1. The flowchart of Semantic Segmentation using FCNs](image)

3.2 Network Configurations

To achieve reasonable segmentation results, we choose VGG16 [12] as the basic network configuration. We replace the original fully-connected layer with convolutional layers and also add several new operations such as deconvolution, fusion and cropping. Input images and corresponding label information are processed as network input and the corresponding output is two-dimensional images which contain each single pixel classification label.
In such network configurations, feature maps after several convolutional and pooling operations become coarse and the sizes are much smaller. To get back to the original image resolutions, deconvolution (upsampling) is applied to the generated feature maps. We will fuse feature maps after pooling. Then a upsampling operation is used on the fusion results to get the segmentation results. Even we can upsampling the previous results of different times and then sum this larger feature maps with previous pooling layer output to generate new feature maps. Then, this new feature maps will be upsampled to produce more smooth and detailed output. These different network configurations are called FCN-32s, FCN-16s and FCN-8s respectively.

4. Experiment

4.1 Data Preprocessing
In the experiment section, all the hand radiographs collected from Li Ka Shing Faculty of Medicine, The University of Hong Kong. We have got 1189 pieces images and saved them as .jpg files. When using DRU classification scheme to evaluate skeletal maturity, physicians only focus on the distal radius and ulna regions. However, the raw data images contain many other regions information such as fingers and black background. These regions will slow down segmentation processing speed and will also decrease the segmentation accuracy. Therefore, a preprocessing step is necessary to extract only the regions of interest (ROI) i.e. the distal radius and ulna regions. After that, all the extracted hand regions will be resized to be an identical resolution. Some extracted DRU region samples are displayed in Figure 2.

4.2 Data Labeling
When training the FCNs and later testing the network performance, we need to label these data. In this experiment, we choose to use LabelMe [13] which is an open source annotation tool. We label the distal radius and ulna regions under physicians’ guidance. In Figure 3, a labeled sample is presented. The ROI is covered by green dot lines and other regions will be considered as background.

4.3 Segmentation Evaluation Metrics
In our experiment, we use two common metrics for semantic segmentation evaluation. The first one is Precision and Recall. Precision reflects the percentage of classified positive samples of all true posi-
tive samples and Recall denotes the fraction of interested samples which have been retrieved over all the data samples. Definitions of Precision and Recall can be written as equation (1) and equation (2)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

Where TP is True Positive, FP is False Positive, and FN is False Negative. The second evaluation metric is called Intersection over Union (IoU). It represents the overlapping rate of segmentation results and ground truth. The definition of IoU can be formulated as equation (3)

\[
\text{IoU} = \frac{A \cap B}{A \cup B}
\]

Where \(A\) and \(B\) represent the segmentation result and ground truth value respectively. Ideally, we want IoU to approach to 1 as much as possible. \(\text{IoU} = 1\) refer to the segmentation results match the ground truth regions perfectly.

5. Result and Analysis

The DRU dataset contains 1189 hand radiographs in total. During the experiment section, 951 images (80%) are randomly selected as training set, and the left 238 images (20%) are used to test trained network performances. All the images are normalized before fed into the proposed network. The normalization operations include mean value subtraction and variation division. To accelerate the network training convergence speed, we choose to use the VGG-16 model which were well trained on ImageNet dataset [14] and finetune it with our DRU dataset. All the models are implemented in Caffe[15] on NVIDIA Quadro M4000. When training deep neural networks, it is easy to occur gradient exploding problems. It is important to initialize learning rate to be a moderate value such as \(1.0 \times 10^{-8}\) in this paper. To overcome loss functions falling into local minima, the momentum value is set to be 0.99, weight decay is \(5 \times 10^{-4}\), the maximal iteration is determined to be 20000 and batch size is 1 during our experiment.

5.1 Network Performance of Different Training Iterations

We use DRU training dataset to train FCN-8s model in various iterations and use our testing dataset to test FCN-8s model. Figure 4 and figure 5 show the radius and ulna pixel-wise accuracy, recall and IoU of the testing dataset.

![Figure 4. Radius Testing Accuracy, Recall and IoU of different iterations](image)

![Figure 5. Ulna Testing Accuracy, Recall and IoU of different iterations](image)
From the above several figures, the total testing accuracy and recall achieve around 98%, IoU is about 97%. The Radius testing accuracy and recall is 97% and radius IoU is around 95%; Ulna only testing accuracy, recall and IoU are 98.5%, 98% and 96.6% respectively.

5.2 Network Performance of Different Configurations

We want to know the influences of DRU semantic segmentation of different network configurations. Therefore, we will test FCN-2s, FCN-4s, FCN-8s, FCN-16s and FCN-32s. In Figure 6 and figure 7, the testing accuracy, recall and IoU are recorded according to different network configurations.

![Figure 6. The Radius Accuracy, Recall and IoU of different network configurations](image1)

![Figure 7. The Ulna Accuracy, Recall and IoU of different network configurations](image2)

From the above figures, we know that FCN-16s and FCN-32s models utilize shallow layer features. They combine the deep features with shallow features and then produce the segmentation results. In the same way, FCN-8s uses more shallow features to assist segmentation performance. Generally speaking, according to the above experiment results, the testing accuracy of FCN-8s is higher than FCN-16s and FCN-32s models. On the other hand, although FCN-4s and FCN-2s involve more shallow features, the network performance has no significant improvement. These two kinds of models have more network parameters and more complicated network structures. After comprehensive considerations, FCN-8s is the best performed network model among all the others for this task.

Finally, the FCN-8s model is used to do DRU semantic segmentation. The results are shown in the Figure 8. The gray region denotes for ulna, while region represents radius, and the black region stands for the image background. The proposed network segment all the 238 testing images in 174 second. Each image only costs 0.73 second to be processed in the experiment.

![Figure 8. Radius and ulna segmentation results](image3)

6. Conclusion

In this paper, we trained a fully convolutional neural network to extract distal of radius and ulna regions of hand X-ray images. The well-trained network can segment bones boundaries subtly in the
experiment. Also, we have compared different network configurations and chosen the one with the best performance i.e. FCN-8s. Additionally, we have found that the proposed network achieved high testing accuracy with a small number of training samples.

Acknowledgments
This work was supported partly by National Natural Science Foundation of China (No. 61872351), Shenzhen Peacock Plan (No.KQJSCX20170331162115349), and Shenzhen Key Basic Research Project (No.JCYJ20180507182506416).

References
[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521 (7553), 436-444.
[2] Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35(5), 1153-1159.
[3] Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging, 35(5), 1240-1251.
[4] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).
[5] Luk, K. D., Saw, L. B., Grozman, S., Cheung, K. M., & Samartzis, D. (2014). Assessment of skeletal maturity in scoliosis patients to determine clinical management: a new classification scheme using distal radius and ulna radiographs. The Spine Journal, 14(2), 315-325.
[6] Kanmitkas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., ...& Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis, 36, 61-78.
[7] Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., & Isgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. IEEE transactions on medical imaging, 35(5), 1252-1261.
[8] Roth, H. R., Lu, L., Farag, A., Shin, H. C., Liu, J., Turkbey, E. B., & Summers, R. M. (2015, October). Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In International on Medical Image Computing and Computer-Assisted Intervention (pp. 556-564). Springer, Cham.
[9] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241). Springer, Cham.
[10] Andrearczyk, V., & Whelan, P. F. (2017). Texture segmentation with fully convolutional networks. arXiv preprint arXiv:1703.05230.
[11] Valipour, S., Siam, M., Jagersand, M., & Ray, N. (2017, March). Recurrent Fully Convolutional Networks for Video Segmentation. In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on (pp. 29-36). IEEE.
[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.15-56.
[13] Torralba, A., Russell, B. C., & Yuen, J. (2010). Labelme: Online image annotation and applications. Proceedings of the IEEE, 98(8), 1467-1484.
[14] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A largescale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.
[15] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678). ACM.