INTERVERTEBRAL DISC DETECTION IN X-RAY IMAGES USING FASTER R-CNN

Ruhan Sa[1], *William Owens*[1], *Raymond Wiegand*[2], *Mark Studin*[3],
Donald Capoferri[4], *Kenneth Barooha*[4], *Alexander Greaux*[4],
Robert Rattray[4], *Adam Hutton* [4], *John Cintineo* [4], *Vipin Chaudhary*[1]
[1]State University of New York (SUNY) at Buffalo
[2]Spine Metrics, Inc.
[3] University of Bridgeport College of Chiropractic
[4] Academy of Chiropractic

ABSTRACT

Automatic identification of specific osseous landmarks on the spinal radiograph can be used to automate calculations for diagnosing ligament instability and injury, which affect 75% of patients injured in motor vehicle accidents. In this work, we propose to use deep learning based object detection method as the first step towards identifying landmark points in lateral lumbar X-ray images. The significant breakthrough of deep learning technology has made it a prevailing choice for perception based applications, however, the lack of large annotated training dataset has brought challenges to utilizing the technology in medical image processing field. In this work, we propose to fine tune a deep network, Faster-RCNN, a state-of-the-art deep detection network in natural image domain, using small annotated clinical datasets. In the experiment we show that, by using only 81 lateral lumbar X-Ray training images, one can achieve much better performance compared to traditional sliding window detection method on hand crafted features. Furthermore, we fine-tuned the network using 974 training images and tested on 108 images, which achieved average precision of 0.905 with average computation time of 3 second per image, which greatly outperformed traditional methods in terms of accuracy and efficiency.

Index Terms— intervertebral disc, detection, deep learning, X-Ray

I. INTRODUCTION

Automatic localization of specific osseous landmarks on the spine radiograph can be used to automate calculations needed for diagnosing ligament instability and other injuries. This serious healthcare condition affects approximately 75% of patients injured in motor vehicle accidents and is a precursor for other spine related diseases. Automation would lead to faster and more accurate diagnosis thereby allowing for faster and more appropriate clinical intervention [1], which leads to better clinical outcome.

Automatic detection of intervertebral discs is an important step towards localizing landmark points in lateral lumbar X-ray images, which provides location information of the discs. This information can then be used for further localizing landmark points within the disc region. Thus the accuracy of the landmark localization depends upon robust intervertebral disc detection results. In this work, we focus on the first step of landmark localizing problem – intervertebral disc detection problem.

Object detection is one of the major research areas in computer vision field due to the challenges brought by low image contrasts, various image scales and etc. Recent work on deep learning technology has made significant breakthrough on perception based applications, which made it a prevailing choice for many visual perception based applications. However training a network from scratch often requires massive amount of training data. For example, state-of-the-art deep learning object detection network Faster-RCNN[2] requires about 150,000 natural images for best performance. The lack of training data has brought a great challenge to medical image applications from utilizing this new technology. In this work, we demonstrate that using very small training dataset, one can achieve great accuracy and increased efficiency by fine-tuning a deep learning network trained on natural images. In particular, we fine-tuned Faster-RCNN network using only 81 lateral lumbar X-Ray images as training data and compared the result with traditional sliding window method on hand crafted features. The average precision rate is increased to 0.65 from 0.03. Furthermore, if trained on a slightly larger dataset of 974 training images, one can achieve 0.905 high average precision rate and the average detection time is greatly reduced to 3 second per image as compared to 82 second per image in traditional methods. This result shows the great potential of using proposed techniques in medical image field when applying deep learning based techniques. In the following sections, different fine-tuning techniques are also discussed in detail.

II. RELATED WORK

Object detection has been a major research area in computer vision field. Traditionally, hand crafted features,

such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) and etc are widely used to train various classifiers. However, compared to recent breakthrough work in deep learning field [3], [4], [5], the performance of traditional methods are far from satisfactory.

Traditionally, object detection is performed through applying trained classifier on image with sliding window fashion or region proposal based fashion. Sliding windows often give more coverage to the image, however it is often too slow for real-world application due to its large redundancy. Much work has been dedicated to accelerate the process [6], [7], [8]. Ren et al proposed Faster-RCNN, where a convolutional neural network is used as region proposal network (RPN)[2]. This technique significantly reduced the object detection time, in the meantime achieved state-of-the-art object detection performance [5]. However, training such a network from scratch requires huge amount of labeled training data, which makes it hard for medical image applications to utilize this new technology, due to the fact that it is extremely hard to collect such a large data with correctly diagnosed labels. In this work, we show that by using very small training data one can achieve satisfactory object detection results by fine-tuning a pre-trained deep learning network.

III. FASTER-RCNN

Faster-RCNN is a state-of-the-art object detection algorithm based on deep learning network. The structure of Faster-RCNN is as shown in Fig 1, where **data** layer is images with various scales, **convolutional neural network** is a regular deep learning classification network, for example, a ResNet. **rois** is the RPN layer, which generates potential object boxes as region of interest. **ROI Pooling layer** takes region of interests and convolutional features as input and generate the bounding box of the objects as well as the corresponding class name. Further details of Faster-RCNN is introduced in [2]. In this work, we fine-tuned this network for intervertebral disc detection task and achieved great performances. We also explored different fine-tuning techniques and compared the performances. More details are discussed in section IV.

IV. EXPERIMENTS

IV-A. Network structure

From Fig 1 we can see that Faster-RCNN requires a convolutional network in **Convolutional Neural Network** layer. Typical image classification deep neural network can be used in this layer. In this work, we used ZF network [9], which is a 5-layered convolutional neural network. We initiate the network parameter with Faster-RCNN ZF network using Caffe [10] for fine-tuning.

IV-B. Evaluation metrics

The detection is judged by average precision based on precision/recall curve. Detection is considered true or false positives based on the area of overlap with ground truth bounding boxes. Correct detection must have overlap α of more than 50% between predicted bounding box B_p and ground truth bounding box B_{gt}, as shown below. Details of the evaluation methods is discussed in [11].

$$\alpha = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p} \tag{1}$$

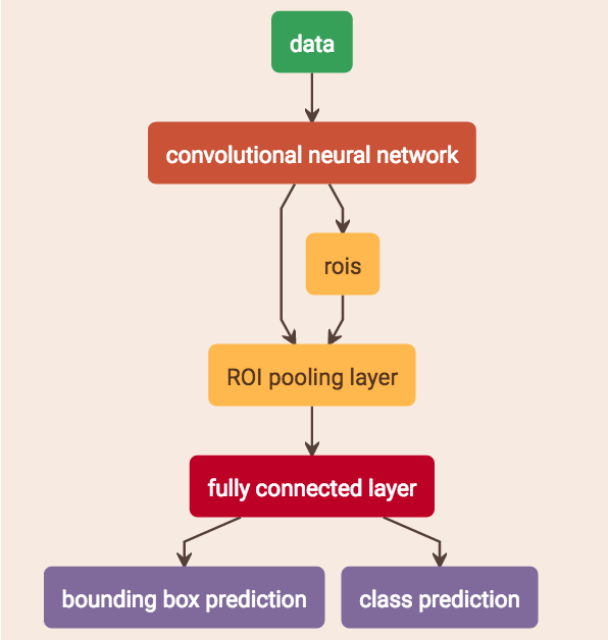

Fig. 1. Simplified structure of Faster-RCNN.

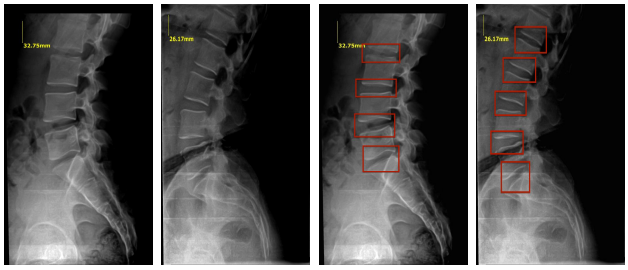

Fig. 2. Example of training data. Left two images show the raw X-ray images and the right two images show the corresponding labeled images. The red rectangle boxes indicate the ground truth location of each visible intervertebral disc.

	shallow tune	deep tune
Average Precision	0.581	0.651

Table I. Average precision comparison between deep tuning and shallow tuning.

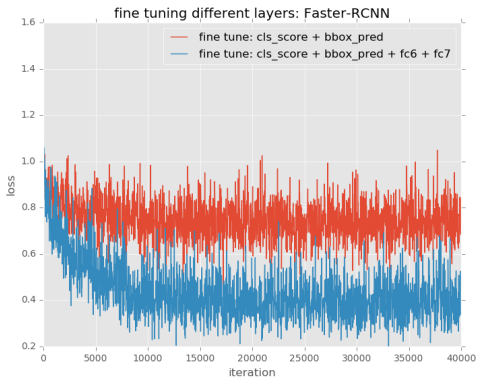

Fig. 3. Effect of shallow and deep fine-tuning on loss change for Faster-RCNN. We can see that deeper fine-tuning yields lower loss convergence.

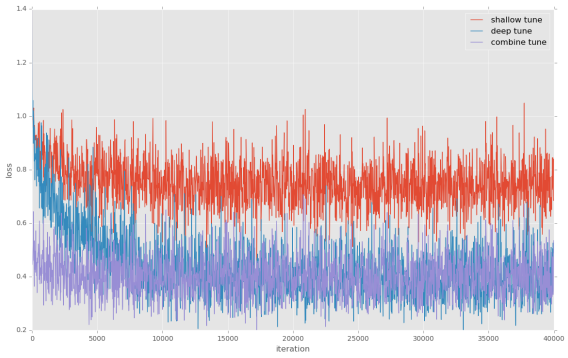

Fig. 4. Effect of shallow and deep fine-tuning VS two-stage training on loss change. We can see that combined tuning does not give lower convergence loss.

IV-C. Training data

Our smaller training dataset consists of 92 lateral lumbar X-Ray images. We randomly divide the dataset into 81 training images and 11 testing images. In the larger dataset, we used 974 training images and 108 testing images. Examples of raw images as well as the labeled training data are shown in Fig 2. As we can see from the image, the size of each intervertebral discs are different. The size of the image also varies from 500x600 to 500x1309 pixel. The number of visible vertebrae in the image varies from 5 to 12 per image.

Using our training data, we fine tune the pre-trained Faster-RCNN ZF network as specified in section IV-A. While fine-tuning, one of the metrics we use to determine the tuning performance is observing the loss change through training iterations. In this work, we used Smooth L1 Loss function, which is described in [12]. Following sections show the performances of different tuning techniques. For the best performance, we used initial learning rate of 0.01, learning

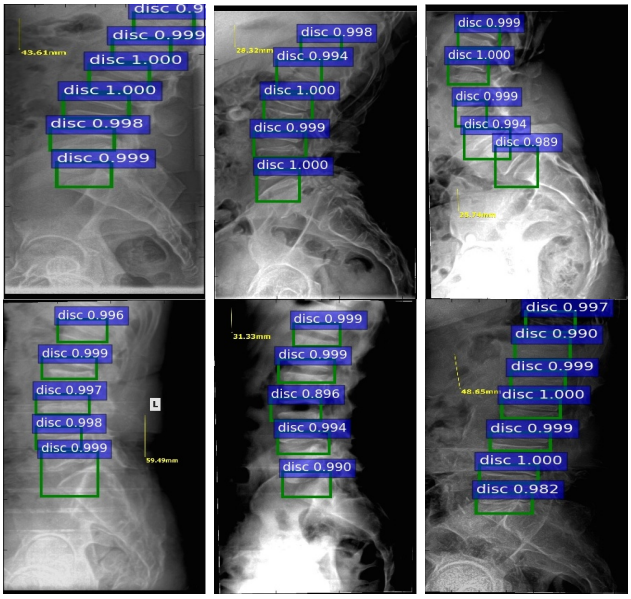

Fig. 5. Example of testing results.

rate dropping ratio is 0.1, weight decay is 0.8, step size for dropping learning rate is set as half of the entire learning iterations.

IV-D. Shallow tuning and deep tuning

In this section, we compare the performances of shallow tuning technique and deep tuning technique. In shallow tuning, we tuned last two layers, which are **cls_score** and **bbox_pred**. In deeper tuning, we tuned last four layers, which are **cls_score, bbox_pred, fc6 and fc7**. Fig. 3 shows that for 40000 training iterations, fine-tuning deeper layer yields better performance. As we can see that deeper layer training converges to a lower loss value by the end of training cycle. In the meantime, Table IV-B shows that deeper fine-tuning also gives better precision performance on testing dataset.

IV-E. Two-stage-tuning

In order to examine if combining shallow and deep fine-tuning together would give performance boost, we fine-tuned network using two-stage-tuning. First we deep tune the network as described in section IV-D and using the trained network, we further train the shallow layer of network, such that the parameters in last layer can have a finer adjustment. From Fig. 4 we see that extra stage of tuning can give faster convergence but achieve similar convergence value.

IV-F. Processing time and qualitative result

The average detection time is 3 second per image on Geforce GTX 1060 mini GPU. Sample testing results are shown in Fig. 5. We can see that the detection result with higher confidence is with great accuracy, so that the number of true positives under high confidence is high.

	AVG precision	AVG time (sec)
HOG+SVM (smaller dataset)	0.032	26
Faster-RCNN (smaller dataset)	0.651	10
HOG+SVM (larger dataset)	0.091	82
Faster-RCNN (larger dataset)	**0.905**	2

Table II. Comparison between traditional sliding window based method and Faster-RCNN based method on both smaller dataset, which contains 81 training image and 11 testing images, and larger dataset, which contain 974 training images and 108 testing images. The average time is the time taken under the corresponding average precision rate.

IV-G. Comparison with baseline traditional method

We compared Faster-RCNN based approach with traditional sliding window approach on both smaller dataset and larger dataset. We extracted Histogram of Oriented Gradient (HOG) features from each training sample and trained SVM classifiers and used sliding window fashion for detection phase. Table IV-G shows the best performing result from each method and we can see that Faster-RCNN based approach outperformed traditional sliding window detection method in both accuracy and efficiency with a large margin in both datasets.

V. CONCLUSION

In this work, we used Faster-RCNN object detection method as the first step towards automatically identifying landmarks from spine X-Ray images. Due to lack of annotated medical images, training deep neural networks can be very challenging. In order to overcome this issue, we show that by fine-tuning a pre-trained deep network on small medical dataset, one can achieve satisfactory results. Our experiments demonstrate that Faster-RCNN based detection method outperformed traditional object detection method by a large margin. Additionally, we also explored different fine-tuning techniques and compared the performances on our medical dataset. For future work, we will experiment on deeper network to improve the detection precision and compare the result with more traditional methods.

ACKNOWLEDGEMENT

This research was based upon work supported by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

VI. REFERENCES

[1] R. Wiegand, N.W. Kettner, D. Brahee, and N. Marquina, "Cervical spine geometry correlated to cervical degenerative disease in a symptomatic group," *Journal of manipulative and physiological therapeutics*, vol. 26, pp. 341–346, 2003.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[6] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014.

[8] Ruhan Sa, William Owens, Raymand Wiegand, and Vipin Chaudhary, "Fast scale-invariant lateral lumbar vertebrae detection andgmentation in x-ray images," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016, pp. 17–20.

[9] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," .

[12] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.