FULLY CONVOLUTIONAL NETWORKS FOR BUILDING AND ROAD EXTRACTION: PRELIMINARY RESULTS

Zilong Zhong[1] ∗, Jonathan Li[1 2] ,Weihong Cui[1 3 4] , Han Jiang[1]

[1] Mobile Mapping Lab, Department of Geography & Environmental Management, University of Waterloo, Ontario N2L 3G1, Canada - (z26zhong, junli, w28cui, h64jiang)@uwaterloo.ca

[2] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, FJ 361005, China – junli@xmu.edu.cn

[3] School of Remote Sensing and Information Engineering, Wuhan University, 430079, China

[4] Collaborative Innovation Center for Geospatial Technology, 430079, China – whcui@whu.edu.cn

ABSTRACT

Available big geoscientific data and modern powerful computation hardware have laid a solid foundation for the prevailing deep learning models in the field of image classification, detection and segmentation. In these models, fully convolutional networks achieve unprecedented success in image segmentation tasks [6]. In this paper, we apply the contemporary image segmentation models in the context of extracting buildings and roads from high spatial resolution imagery. We estimate the influence of filter stride, learning rate, input data size, training epoch and fine-tuning on model performance. Selected Massachusetts road and building datasets are used for training, validation, and testing the performance of the models with different parameters. As a result of combining shallow fine-grained pooling layer outputs with the deep final-score layer or abandoning coarse-grained pooling layers, the extraction precision rate of the best modified model improves significantly to over 78%.

Index Terms**—** Deep learning, fully convolutional networks, object extraction, high spatial resolution imagery

1. INTRODUCTION

With the rapid development of earth observation and remote sensing technologies, satellites and aircrafts have acquired various kinds and enormous amounts of high spatial resolution data. However, sheltering, overlapping, disordering, interlacing and shadows exist in remotely sensed imagery. In addition, both inter-class and intra-class variances are prevalent in high spatial resolution imagery [1-3]. However, municipal management officers, traffic and transportation practitioners, urban planners, and geodetic academic researchers and geographic information system (GIS) users have difficulty extracting valuable information from the chaotic, complicated and non-intuitive data. Moreover, topological maps available on the Open Street Map website through crowd sourcing projects provide much clearer and cleaner products, which ordinary people can easily access. Thus, extracting semantic information from high spatial resolution aerial imagery plays an important role in automatic simplifying, digitalizing and visualizing the high-dimension, high-variation and high-resolution imagery. Fortunately, big-data-driven and hardware-based deep learning models have demonstrated an enormous ability to extract features in the domain of image classification, detection and segmentation [4-6]. Because remote sensing data, such as multi-spectral, high-spectral, high spatial resolution imagery and lidar point clouds, are inherently and naturally big data, we intuitively resort to deep learning models to automatically learn hierarchical features and extract semantic information from the raw data.

Deep convolutional neural networks (CNNs), which have achieved unprecedented superb results in the computer vision and pattern recognition contests, have been extensively studied in previous research [7-10]. Recently, with the prevalence of CNNs and the availability of big remote sensing data, more and more articles have drawn on the feasible automated feature generation ability of deep learning models and applied them to the field of high spatial resolution imagery processing. In [7], simple CNN models were adopted in object labelling on aerial imagery and combined prior knowledge with post-processing steps to mitigate omission noise and registration noise [8]. Specifically, multi-layer neural networks were employed as a post processing step that assist in refining the outputs of CNNs. In [4], random forest classifiers were used to label pixels based on features learned from CNNs and manually designed feature extractors. Afterwards, a conditional random field (CRF) smoothed the previous coarse and semantic prediction [7]. However, the models used to learn features were still composed of typical CNN layers, which are constrained by the objective of obtaining high level information instead of local details.

In this paper, we experimentally apply the latest image segmentation CNN models, which combine high order semantic meaning with low level fine-grained appearance, to the task of object extraction from high spatial resolution imagery. Furthermore, we test and analyse the different hyper parameters, learning strategies and input image sizes to determine the capability of fully convolutional networks in the context of large scale, high resolution and widely noisy aerial imagery.

2. RELATED WORK

From one perspective, extracting objects from aerial imagery can be treated as a classification problem, where each pixel is classified in the imagery into certain categories. Pixel-based, object-based and neural network classification methods are three common categories of approach adopted in many articles [11-13]. Four different main methodologies were reviewed in [11] to analyse and classify urban impervious surfaces and summarize the data analysis and information processing emerging trends. In [12], the object-based random forests classifiers, which employed natural block road boundaries as prior knowledge, were trained for land cover mapping and urban structure type classification.

However, object-based classification models are highly dependent on intermediate segmentation results. If the segmented parcels appear to be poor, the object-based classification frameworks will generate low quality outcomes correspondingly. In contrast, neural network classifiers integrate the advantage of pixel-based and object-based methods. As the input data is always uniform square patches, instead of pixels or objects, the neural network methods are also known as the patch-based methods. In this paper, patch-based learning methods that combine the coarse high-level semantic information and shallow low-level fine details are applied to the extraction of roads and buildings.

From another point of view, extracting objects from imagery equals to finding the objects of interest and separating them out of the backgrounds. Therefore, extraction of impervious surfaces, such as rooftops and roads, in high resolution imagery can be regarded as semantic segmentation tasks. In late 2014, Long innovatively transferred the deep learning models for image classification by convolutionizing the top fully connected layers and presented multi-scale and fully convolutional neural networks (FCNs), which achieved a striking improvement over previous methods in the Pascal VOC 2012 image segmentation benchmark [6]. Based on FCNs, in early 2015, Zheng et al. proposed a CRF-RNN model that integrated the feature learning properties of fully convolutional networks with the smooth confine enforced by condition random fields (CRFs), which obtained the state-of-the-art results [9]. Half a year later, Liu et al. combined the deep parsing network (DPN) that bond Markov Random Field (MRF) and CNNs, and reduced the training time significantly by approximating a single parameter learning iteration [10]. In addition, based on FCNs, Lin et al. conferred a message passing inference, which not only drastically increased the efficiency of the forward inference of the FCNs and CRFs to learn the structural information of imagery, but achieved the highest accuracy in the Pascal VOC 2012 contest [5].

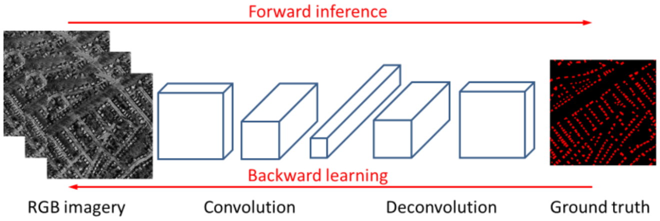

Fig. 1. Fully convolutional networks

Considering the latest image segmentation and pattern recognition advances, especially the FCNs, it is intuitive to apply the deep convolutional models to the field of high spatial resolution image processing. Unlike the original fully convolutional models, we extend the model with extra fine-grained pooling layer outputs combined in the final scoring layer and train new models using different learning rates, input sizes and fine-tuned models.

3. FULLY CONVOLUTIONAL NETWORKS

As seen in Section 2, most of the latest top image semantic segmentation models are based on the FCNs. Actually, FCNs are derived from the large scale image classification model VGG [6]. Compared to large scale image classification models, the FCNs can not only predict what the object of interest is, but also conjecture the spatial location and contour details, which are vital to semantic segmentation.

The convolutional layer in the FCNs can be formulated using the equation below.

$$y_{ij} = f[\sum_{m_i,m_j=0}^{k} g(x_{s_i+m_i,s_j+m_j})] \qquad (1)$$

In equation (1), y_{ij} represents the output of the convolutional layer in i row and j column, x represents the input pixels from the lower layer, k is the size of the convolutional filter, and s is the sampling stride. The symbol $g(\bullet)$ within the summation sign denotes the convolution function over a designated area, and $f(\bullet)$ denotes the output normalized function.

The FCN model has been visualized in **Fig. 1**, where the input is a the 3×500×500 optical aerial image. At the forward inference period, the three-channel high resolution data was down-sampled in the low convolutional layers with relatively small convolutional filters, and up-sampled in the high deconvolutional layers with corresponding filters. Stochastic gradient descent was used to calculate the backward learning parameters according to the difference between model predictions and ground truth images, which are the same size as the input data.

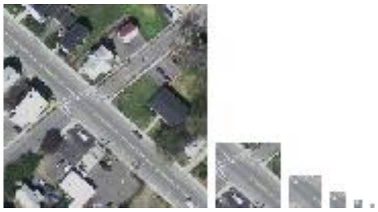

Fig. 2. Different sampling stride (from left to right: training image, 32s, 16s, 8s, 4s and 2s filters)

Because the PASCAL image segmentation datasets present enormous differences from the road and building extraction datasets, we added the pre-trained FCN model with a new 4-stride pooling layer output into the final score layer and fine-tuned the model with new high spatial resolution data. As illustrated in **Fig. 2**, as the stride of the filters becoming smaller, the models have increasingly detailed spatial and textural information to learn from.

Theoretically, the FCN's computation consumption could be much higher than that of the ordinary object recognition models. Nevertheless, in the implementation process of FCNs, the large portion of overlapping in the receptive fields notably enhance the forward inference and backward learning process. Particularly, the top fully connected layers of FCNs make a conversion to convolutional layers, using convolutional kernels with the same size of the data blob from lower layers.

4. RESULTS AND DISCUSSION

The road and building detection datasets, published by the University of Toronto, include Massachusetts' road dataset and building dataset [7]. The resolution of the images in the Massachusetts datasets is 1 meter and each image consists of 3×1500×1500 pixels. Specifically, the road dataset contains 1,711 aerial images, which cover more than 2,600 km^2 in total. Also, the building dataset has 151 aerial images of about 340 km^2 in the Boston area. Both of them can be downloaded from internet. Since some aerial images used for training and validate include blank areas where the corresponding ground truth data fail to match up, these disabled data are eliminated from the dataset to attain reasonable building and road extraction capability. After the qualitative-based selection of the images, the size of the updated road datasets is 6.48 GB and the building counterpart holds 1.48 GB.

To make full use of the existing pre-trained models, we divide each original image into nine uniformed 3×500×500-pixel images, which make our data nine times larger. Then, the two datasets are respectively separated into training datasets and validation datasets. To some extent, the following experiments exhibit the likely optimum approaches to manipulate the FCNs in extracting thematic information from high resolution aerial imagery. Precision, recall and intersection over union (IU) accuracies are adopted to estimate the prediction performance of FCN models.

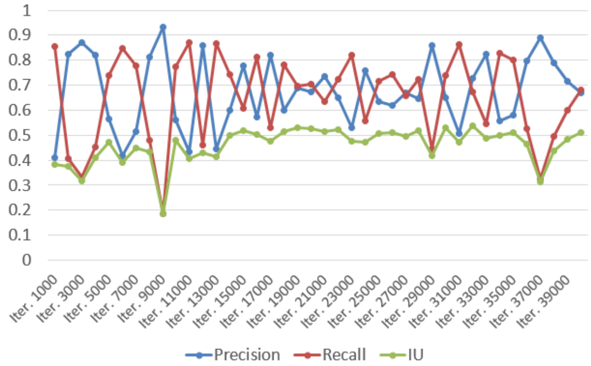

Fig. 3. Test precision, recall and IU accuracies on the pruned FCN-2s network.

Table 1. Performance comparison of different models

Model	Object	Prec.	Rec.	IU	Iter.
FCN-8s	Road	0.43	0.49	0.30	10000
	Building	0.54	0.67	0.43	20000
Enlarged FCN-4s	Road	**0.71**	**0.66**	**0.52**	4000
	Building	0.73	0.60	0.50	12000
Pruned FCN-2s	Road	0.58	0.61	0.42	39000
	Building	**0.78**	**0.61**	**0.52**	15000

(*Prec.: precision accuracy; Rec.: recall accuracy; IU: intersection of unit; Iter.: optimum iteration*)

4.1. Fine-tuning FCN-8s directly

At first, we directly fine-tuned the new model based on the pre-trained FCN-16s-PASCAL model. The learning rate is set to 1×e^{-14}, and the model is trained for 20,000 iterations.

Then, FCN-8s building extraction models were fine-tuned based on the previous intermediate building extraction model and the new trained models went through another 30,000-iteration training. The learning rate was set the same as the previous step. Meanwhile, new road extraction models, which derive from the same trained model, were fine-tuned with the learning rate of 1×e^{-13}.

4.2. Fine-tuning extended FCN-4s

In response to the preceding fine-tuned experiments with unsatisfying outcomes, we added the 4s pooling output to the final score layer, and fine-tuned the road and building extraction FCN-4s network directly on FCN-16s model with the fixed learning rate 1×e^{-14}. According to **Table 1**, an obvious increase in accuracy can be attained after extending the previous training model. Especially, the new trained road extraction model achieved 71% precision rate and 66% recall rate in the 4000th-iteration.

4.3. Fine-tuning pruned FCN-2s

As present in **Fig. 3**, the final pruned building extraction FCN-2s models, which discard 32s, 16s and 8s pooling layers

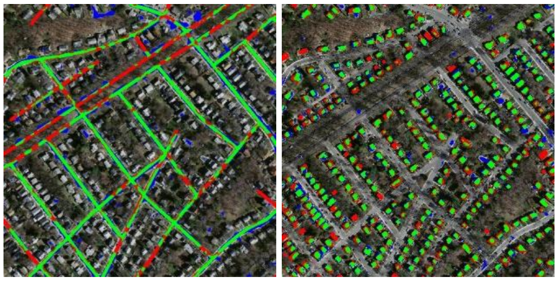
(a) (b)

Fig. 4. Extraction results: (a) road extraction outcomes of 4,000-iteration enlarged FCN-4s model; (b) building extraction results of 15,000-iteration pruned FCN-2s model.

and encompass more fine-grained 2s pooling layer output on top of the final scoring layer, exhibited a steady state between 15,000th-iteration and 30,000th-iteration. They were fine-tuned for 40,000 iterations based on the pre-trained FCN-16s model with the fixed learning rate of $1\times e^{-13}$. As **Table 1** shows, the pruned FCN-2s network achieved a 60% recall rate and a 73% precision rate. On the contrary, with the same parameter settings, the FCN-2s road extraction models performed worse than their FCN-4s counterparts. This phenomenon demonstrated that road extraction tasks require low-level and texture information more than high-order and semantic expression.

The best results of road and building extraction are shown in **Fig. 4**, in which green parts represent the true positive extraction pixels, the red areas denote the false negative labelling pixels, and the blue parts are false positive pixel labels. In view of the outcomes obtained from limited modifications and training strategies, the enlarged and pruned models appear more competitive than the directly used ones. Thus, the results are encouraging and promising.

4.4. Different input sizes

For the purpose of comparison, we also built a different aerial imagery dataset with each image having 3×100×100 pixels. However, the results of experiments based on directly fine-tuned networks did not present improvement in accuracy. Nevertheless, further experiments are necessary to exploit the influence of input data sizes.

5. CONCLUSIONS

This paper has presented preliminary outcomes of different fine-tuned networks derived from the pre-trained FCN-16s model. According to the results of the comparison experiments, it is not feasible to directly transfer the existing FCN pre-trained models into tasks of road and building extraction. However, adding outputs from pooling layers with smaller sampling stride into the final score layer have improved the labelling accuracy in the road extraction scenarios significantly. Furthermore, discarding redundant and large pooling layers have benefited the classification accuracy of building extraction. These results imply the orientation of future research on this topic. Interestingly, the pruned FCN-2s road extraction models have performed worse than the enlarged FCN-4s network. The reason lies in inherent representation properties of different pooling layers. Moreover, learning rates have presented an obvious impact on training procedures. When training with higher learning rates, the new trained FCN models have shown higher fluctuation in extraction and are unlikely to converge. Future research will focus on simpler network structures and high-efficiency training strategies to further interpret this issue.

Acknowledgement: This work is supported by the China Scholarship Council (CSC).

6. REFERENCES

[1] A. Katartzis, H. Sahli, “A stochastic framework for the identification of building rooftops using a single remote sensing image,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 1, pp. 259-271, Janurary 2008.

[2] S. Agarwal, L.S. Vailshery, M. Jaganmohan, H. Nagendra, “Mapping urban tree species using very high resolution satellite imagery: comparing pixel-based and object-based approaches,” *ISPRS International Journal of Geo-Information*, vol. 2, no. 1, pp. 220-236, March 2013.

[3] S.R. Garrity, C.D. Allen, S.P. Brumby, C. Gangodagamage, N.G. McDowell, D.M. Cai, “Quantifying tree mortality in a mixed species woodland using multitemporal high spatial resolution satellite imagery,” *Remote Sensing of Environment*, vol. 129, pp.54-65, February 2013.

[4] S. Paisitkriangkrai, J. Sherrah, P. Janney, A. Hengel, “Effective semantic pixel labelling with convolutional networks and conditional random fields,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36-43.

[5] G. Lin, C. Shen, I. Reid, A. Hengel, “Deeply learning the messages in message passing inference,” In *Advances in Neural Information Processing Systems*, 2015, pp. 361-369.

[6] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” 2014, *arXiv preprint arXiv:1411.4038*.

[7] V. Mnih, “Machine learning for aerial image labeling,” PhD diss., University of Toronto, 2013, 109p.

[8] Y. Shu, “Deep Convolutional Neural Networks for Object Extraction from High Spatial Resolution Remotely Sensed Imagery.” PhD diss., University of Waterloo, 2014, 133p.

[9] S. Zheng, et al., “Conditional random fields as recurrent neural networks,” 2015, *arXiv preprint arXiv*: 1502.03240.

[10] Z. Liu, et al., “Semantic Image Segmentation via Deep Parsing Network,” 2015, *arXiv preprint arXiv*: 1509.02634.

[11] Q. Weng, “Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends,” *Remote Sensing of Environment*, vol. 117, pp. 34-49, February 2012.

[12] M. Voltersen, C. Berger, S. Hese, C. Schmullius, “Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level,” *Remote Sensing of Environment*, vol. 154, pp. 192-201, November 2014.

[13] D. Duro, S. Franklin, M. Dubé, “A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,” *Remote Sensing of Environment*, vol. 118, pp. 259-272, March 2012 .