SRC: Transferring scale-independent features to support multi-scale object recognition with deep convolutional neural network

Xiran Zhou
School of Geographical Sciences & Urban Planning, Arizona State University
Tempe, AZ, U.S.
xrzhou@asu.edu

1 Introduction

Object detection is an essential remote sensing data application that attempts to detect and recognize semantically meaningful instances from complicated context and background [1]. Recently, a great number of works have reported that the state-of-the-art convolutional neural networks (CNNs), significantly facilitates automatically detecting a predefined class of object from high resolution aerial or satellite imageries [2]. However, different from the object in a photo, the objects in remote sensing imageries are always represented by largely varied scale or spatial resolution. Scale has become a critical attribute for recognizing an object from aerial or satellite imagery [3]. Up to now, the state-of-the-art DCNNs for object detection requires predefined scale, which produces a big challenge to automatically adjust the representative scale for an object on the ground [3-4].

Transfer learning is a branch of machine learning techniques that focuses on reusing the exiting data, knowledge or models to solve a new research problem [8]. Deep learning creates independent CNNs for different object detection tasks. Otherwise, transfer learning only builds one CNN, and attempts to transfer the features gained from this CNN to support to detect the small-scale airplane. It is agreed that fine-tuning the deep architecture of a CNN and collect the needed data are computationally-intensive and time-consuming. Transfer learning enables a machine to transfer the already gained knowledge to solve new tasks in an ad-hoc manner, and supports to conduct zero-shot and one-shot learning in an application with small-scale dataset.

A solution to effectively deal with cross-scale data features is important for successfully implementing CNNs to conduct efficient object detection from remote sensing imageries [1]. This paper focuses on transferring the pretrained knowledge derived from large scale object, to support to detect other objects represented within different scales. Based on the architecture of Faster R-CNN, this paper reports a new DCNNs for conducting scale-independent object detection from remote sensing imagery.

2 Methodology

The proposed CNN for object detection depends on the architecture of Faster R-CNN [6], which effectively supports to discover the representative region proposals through integrating the results from the Region Proposal Network (RPN) and those from the classification layer. This paper proposes a new algorithm called atrous region proposals to promote the efficiency of multi-scale region proposal searching, and exploits atrous convolution in the feature extraction layer.

2.1 Atrous convolution

Although pooling layer supports to reduce computational load and facilitates extracting rotational-invariance features, it may fail to support to automatically adjust the field-of-view of a filter, and always leads to the loss of spatial details being beneficial to the representation of an object.

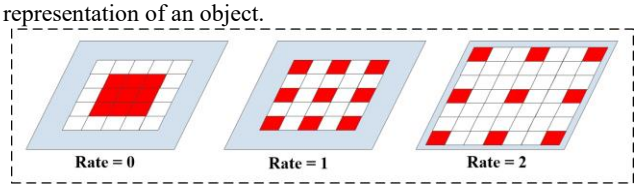

Figure 1: Illustration on atrous convolution layer.

Atrous convolution supports to filter the image through altering the size of the receptive field view in generating region proposal, to control the loss of spatial details [7]. Figure 1 illustrates the principle of atrous convolution. A pixel is filtered based on its neighboring pixels over different distances (rates). Then, the filtering results within a variety of distances are integrated. Due to page limit, more details about atrous convolution can be read in Reference [7].

2.2 Atrous region proposals

This paper aims to extending the receptive field of RPN, which is a critical element used for bounding box searching with optimized dimension and shape. In the Faster R-CNN model, the dimension and the shape of region proposals are controlled by two parameters: archer size and ratio. However, the scales of the same class of objects might be significantly varied in remote sensing imageries [2]. This makes the range of archer size and ratio vary largely to characterize the objects represented by various scales, which posing challenging for object localization due to instability of representative region proposals search.

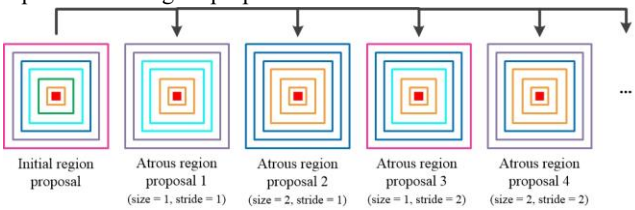

Figure 2: Illustration on atrous region proposals.

To address the challenge mentioned above, applying the idea of atrous convolution, this paper creates a new region proposal called

atrous region proposal. Figure 2 shows the structure of atrous region proposals. Red box denotes a a pixel. Orange, green, cyan, blue, purple and pink line respectively denotes this pixel's neighboring pixels over multiple distances. Pixels in the red box and its surrounding lines constitutes an object. Based on the initial region proposal generated by Faster R-CNN, the proposed method creates a variety of atrous region proposals with the neighboring pixels over multiple distances. The atrous region proposals shown in Figure 2 selects the center pixel-oriented partial context to create a small-scale region proposal.

The dimension of atrous region proposal is controlled by size and stride. Size defines the number of neighboring pixels to be replaced. Stride controls the way how replacing the neighboring pixels. For example, when size equals 1, only the green line is replaced by the orange one in atrous region proposal 1. When size equals 2, the green and cyan line are replaced by the orange one in atrous region proposal 2. Additionally, stride is used to control how replacing around the neighboring pixels. When stride equals 1, after replacing the green line with the orange one, the cyan line is used to replace the blue one in atrous region proposal 1. In atrous region proposal 3, when stride equals to 2, after replacing the green line with the orange one, the proposed method jumps the cyan line and uses the blue line to replace the purple one.

3 Experiment

The dataset for evaluating the proposed DCNN for object detection is selected from a large-scale benchmark dataset called CSRS-SIAT [8]. SIAM-CSRS includes around 70 scenic and object categories of satellite imageries over scales, and each category contains 1000 images. To our knowledge, CSRS-SIAT is the only benchmark dataset that provides cross-scales imageries for every scene class. In this paper, we select dam, airplane, oil tank, football field, island, swimming pool, lake/pond and crater. 200 samples of every category were labeled, among them 120 samples have a similar scale, and other 80 samples have different scales. Figure 3 shows the selected samples of these eight categories. Red box denotes the labeled bounding box for training and test.

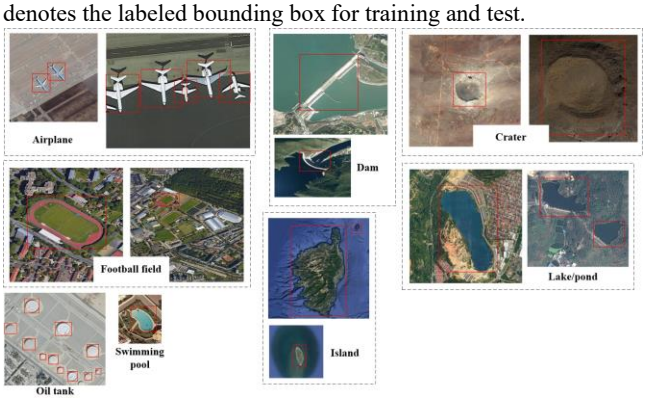

Figure 3: Illustration on experimental datasets.

The first experiment compares the result generated by similar scale object detection with the classic Faster RCNN and the proposed method. Another one compares the result of multi-scale object detection with the classic Faster RCNN and the proposed method. For each experiment, 80% samples were randomly selected for training, and the rest were used for evaluation.

Table 1: Comparison of object detection results.

	Similar scale		Cross-scale	
Category	Faster R-CNN [2]	Proposed method	Faster R-CNN [2]	Proposed method
Dam	90.00%	92.5%	81.5%	86.00%
Airplane	95.83%	97.5%	86.00%	90.00%
Oil tank	98.33%	98.33%	94.00%	95.00%
Football field	87.5%	90.83%	79.50%	82.00%
Island	100%	100%	93.00%	96.00%
Swimming pool	100%	100%	—	—
Lake/pond	90.83%	92.50%	86.00%	89.00%
Crater	91.67%	93.33%	81.00%	84.00%

Table 1 lists average precision (AP) results of object detection. AP is a commonly-used evaluation measure to assess the accuracy of CNN for object recognition [2]. The precision ranges from 87.5% to 100% due to the characteristics of object, the complexness of context, and the land cover where an object locates in. Faster R-CNN and the proposed method works well on similar scale object detection, and the proposed method slightly outperforms the classic Faster R-CNN. This is because atrous convolution performs better than pooling layer in features extraction [9]. Additionally, the results of similar-scale object detection prove that the atrous region proposals proposed remains the efficiency of object detection.

When the objects are represented by different scales, significant decreases of AP are observed in the results by using Faster R-CNN. This shows that it is challenging for Faster R-CNN to predefine appropriate archer size for precise object localization if the scale of objects varies significantly. Through creating extensive region proposals, the proposed method outperforms Faster R-CNN to a certain extent. The work above shows that the RPNs in Faster R-CNN can effectively support multi-scale object detection with modification considering spatial shape and scale.

REFERENCES

[1] G. Cheng and J. Han. (2016). A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 117, 11-28.

[2] H. Wu and Z.L. Li (2009). Scale issues in remote sensing: A review on analysis, processing and modeling. Sensors, 9(3), 1768-1793.

[3] J. Li, X. Liang, S. Shen, et al. (2018). Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia, 20(4), 985-996.

[4] T. Y. Lin, P. Dollár, R. B. Girshick, et al. (2017). Feature Pyramid Networks for Object Detection. In IEEE Conference on Computer Vision and Pattern Recognition, 1(2), 4.

[5] S. J. Pan and Q. Yang (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[6] S. Ren, K. He, R. Girshick, et al. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 1137-1149.

[7] L.C. Chen, G. Papandreou, I. Kokkinos, et al. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.

[8] Y. Shen, X Zhou, J. Liu, et al. (2018). CSRS-SIAT: A benchmark remote sensing dataset to semantic-enabled and cross-scales scene recognition. In IEEE International Geoscience and Remote Sensing Symposium, 2018.