Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Multi-scale object detection in remote sensing imagery with convolutional neural networks

Zhipeng Deng [a], Hao Sun [a], Shilin Zhou [a,*], Juanping Zhao [b], Lin Lei [a], Huanxin Zou [a]

[a] *College of Electronic Science, National University of Defense Technology, Changsha, China*
[b] *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China*

ARTICLE INFO

Article history:
Received 31 May 2017
Received in revised form 19 March 2018
Accepted 16 April 2018
Available online xxxx

Keywords:
Object detection
Deep learning
Convolutional neural networks
Multi-modal remote sensing images

ABSTRACT

Automatic detection of multi-class objects in remote sensing images is a fundamental but challenging problem faced for remote sensing image analysis. Traditional methods are based on hand-crafted or shallow-learning-based features with limited representation power. Recently, deep learning algorithms, especially Faster region based convolutional neural networks (FRCN), has shown their much stronger detection power in computer vision field. However, several challenges limit the applications of FRCN in multi-class objects detection from remote sensing images: (1) Objects often appear at very different scales in remote sensing images, and FRCN with a fixed receptive field cannot match the scale variability of different objects; (2) Objects in large-scale remote sensing images are relatively small in size and densely peaked, and FRCN has poor localization performance with small objects; (3) Manual annotation is generally expensive and the available manual annotation of objects for training FRCN are not sufficient in number. To address these problems, this paper proposes a unified and effective method for simultaneously detecting multi-class objects in remote sensing images with large scales variability. Firstly, we redesign the feature extractor by adopting Concatenated ReLU and Inception module, which can increases the variety of receptive field size. Then, the detection is preformed by two sub-networks: a multi-scale object proposal network (MS-OPN) for object-like region generation from several intermediate layers, whose receptive fields match different object scales, and an accurate object detection network (AODN) for object detection based on fused feature maps, which combines several feature maps that enables small and densely packed objects to produce stronger response. For large-scale remote sensing images with limited manual annotations, we use cropped image blocks for training and augment them with re-scalings and rotations. The quantitative comparison results on the challenging NWPU VHR-10 data set, aircraft data set, Aerial-Vehicle data set and SAR-Ship data set show that our method is more accurate than existing algorithms and is effective for multi-modal remote sensing images.

1. Introduction

Object detection in very high resolution (VHR) remote sensing images is to determine if a given aerial or satellite image contains one or more objects belonging to the class of interest and locate the position of each predicted object in the image (Cheng and Han, 2016). The term 'object' used in this paper mainly refers to man-made objects (e.g. aircrafts, vehicles, storage tanks and ships, etc.) that have sharp boundaries and are independent of background. As a fundamental problem faced for remote sensing image analysis, object detection in remote sensing images plays an important role for both military and civilian applications. However, it's still a challenging problem due to the varying visual appearance of objects caused by occlusion, illumination, shadow, viewpoint variation, resolution, polarization, speckle noise, etc. Furthermore, the explosive growth of remote sensing images in quantity and quality creates an extremely high computational costs, which also increases the difficulties of object detection for near-real-time applications.

Over the last decades, numerous detectors has been developed for detecting different types of objects in remote sensing images. Cheng and Han (2016) reviewed various object detection algorithms in optical remote sensing images and categorized them into four groups, namely template matching-based methods,

* Corresponding author.
E-mail addresses: zpdeng@whu.edu.cn (Z. Deng), hsun@nudt.edu.cn (H. Sun), slzhou@nudt.edu.cn (S. Zhou), juanpingzhao@sjtu.edu.cn (J. Zhao), llei@nudt.edu.cn (L. Lei), hzzou@nudt.edu.cn (H. Zou).

https://doi.org/10.1016/j.isprsjprs.2018.04.003
0924-2716/© 2018 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

Please cite this article in press as: Deng, Z., et al. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.04.003

knowledge-based methods, object-based image analysis-based methods, and machine learning-based methods. For synthetic aperture radar (SAR) images, El-Darymli et al. (2013) categorized the object detection methods into three major taxa: namely single-feature-based methods, multi-feature-based methods, and expert-system-oriented methods. In recent years, due to the advance of the machine learning technique, particularly the deep learning based models with powerful feature representations, many approaches consider object detection as a region-of-interest (RoI) classification problem using deep features and have shown more impressive success for certain object detection tasks than hand-craft features based detectors (Cheng et al., 2016; Girshick et al., 2016; Ren et al., 2015; Deng et al., 2017; Tang et al., 2016). In these approaches, object detection processing is split into two distinctive stages: proposal generation and object classification.

The proposal generation stage aims to generate bounding boxes of object-like targets. The most common paradigm is based on a sliding-window search in which each image is scanned in all positions with different scales. While real-time detectors are available for specific classes of objects, e.g. Constant False Alarm Rate (CFAR) based ship detector in SAR images (Kuttikkad and Chellappa, 1994) or Aggregated Channel Features (ACF) feature based vehicle detector in aerial images (Liu and Mattyus, 2015), it has proven difficult to design multi-class detectors under this paradigm. Furthermore, searching for objects in high-resolution broad-area remote-sensing images led to heavy computational costs. Another popular paradigm samples hundreds of object-like regions, using a visual saliency attention stage, to reduce the search space for the whole image. Although successful detectors are available for multi-classes of objects, e.g. ten-class VHR object detection (Cheng et al., 2016) or simultaneous airport detection and aircraft recognition (Zhang and Zhang, 2016), the RoI generation capability of those saliency analysis based methods may become limited or even impoverished under complex background. What's more, these methods are computationally expensive, at 3 s per image (about 600 × 800 pixels) in a CPU implementation.

The object classification stage infers each region's category by learning a classifier. As object-like region is usually carried out from feature space, powerful feature representation is very important for constructing a high-performance object detector. Convolutional neural networks (CNNs) are among the most prevalent deep learning methods (Dean et al., 2012; Yao et al., 2016; Cheng et al., 2017; Zhang et al., 2016b; Yuan et al., 2015; Feng et al., 2016), which can be employed as an universal feature extractor. Compared with hand-crafted features or shallow-learning-based features, CNN features relying on neural networks of deep architecture are more powerful for representation, which can significantly improve the performance of object detection (Cheng et al., 2016; Girshick et al., 2016). However, all candidate regions should be cropped and scaled into a fixed size (e.g. 224 × 224) supported by the CNN. This pre-requisite warping technique may reduce some critical and discriminative properties for object categories with larger size, resulting in low detection accuracy.

While the aforementioned deep learning based object detection methods have shown impressive success for some specific object detection tasks in remote sensing images, they are all trained in clumsy and separate multi-stage pipelines. On the one hand, how to extract good potential object-like regions is a critical task and computational bottleneck for accurate object detection in large-scale remote sensing images. The existed visual saliency attention based region generation methods involve abundant human ingenuity for features design that is only effective for specific classes of object detection task. On the other hand, region classification ignores the fact that the frame of detection task is a regression problem. In addition, less concern has been given to multi-class object detection tasks, whereas automatically identifying multi-class objects simultaneously plays a significant role for the intelligent interpretation of remote sensing images. Therefore, a good detection model should be able to unify the above two distinctive stages into one unified framework that can detect multi-class objects simultaneously, and be applicable to the variety of data sources.

In the field of computer vision, object detection is one of the most fundamental and challenging problems. In 2013, a breakthrough was made by Girshick et al. (2016), who proposed region-based CNN (R-CNN) detector that improves mean average precision (mAP) by more than 50% relative to the previous best result (Felzenszwalb et al., 2010). Since then, considerable efforts have been made to improve the detector along the R-CNN based pipeline. The most successful improved detector is Faster R-CNN (Ren et al., 2015), which consists of a region proposal network (RPN) for predicting candidate regions, and an object detection network (Girshick, 2015) for classifying object proposals and refining their spatial locations. It is an end-to-end data-driven detector that takes an image as input and outputs the location and category of objects simultaneously, which can effectively overcome the aforementioned drawbacks of the existing deep leaning based object detection methods in remote sensing images. This has motivated us to move from the multi-stage pipelines to the era of unified detection framework.

However, although the Faster R-CNN based methods have proven to be very successful for detecting objects such as cars, people, or dogs in nature scene images, they are not specially designed to detect small objects in large images, several challenges in remote sensing images limit their applications in earth observation community:

(1) Objects often appear at very different scales in remote sensing images. On the one hand, the scale variability is caused by image resolution. On the other hand, different object categories have large size differences. As shown in Fig. 1, ship and bridge have large difference in size. However, Faster R-CNN generates candidate object-like regions by sliding a fixed set of filters with a single receptive field over the topmost convolutional feature maps, which creates an inconsistency between the size variability and fixed filter receptive fields. As shown in Fig. 1, a fixed receptive field (illustrated in the shaded area) cannot match the scale variability of different objects in remote sensing images. For small size or large size objects, the detection performance tends to be particularly poor.

(2) Objects in large-scale remote sensing images are relatively small in size and appear in densely distributed groups (like the storage tanks in Fig. 1). Whereas Faster R-CNN struggles with small objects, this is because CNN feature used for detection is pooled from the topmost convolutional feature map with lower resolution. After multiple downsampling, the object size in the topmost convolutional feature map is 1/16 of the original size in the input images. This may lose some important information for small objects and lead to some missing detections.

(3) Remote sensing images are enormous (often hundreds of megapixels), and there's a relative dearth of labeled training data. The available manual annotation of objects are not sufficient in number for properly training the Faster R-CNN based methods. The collection of labeled samples through photointerpretation or terrestrial campaigns is time consuming and expensive, often requires expertise background.

(4) Remote sensing images are generated by different instruments (e.g., multi/hyperspectral, SAR, etc.) with different resolutions. Remarkable efforts have been made in develop-

Please cite this article in press as: Deng, Z., et al. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.04.003