ISPRS Journal of Photogrammetry and Remote Sensing 146 (2018) 182–196

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images

Yansheng Li[a], Yongjun Zhang[a,*], Xin Huang[a], Alan L. Yuille[b]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

ARTICLE INFO

Keywords:
Multi-class geospatial object detection
Deep networks
Scene-level supervision
Discriminative convolutional weights
Class-specific activation weights

ABSTRACT

Due to its many applications, multi-class geospatial object detection has attracted increasing research interest in recent years. In the literature, existing methods highly depend on costly bounding box annotations. Based on the observation that scene-level tags provide important cues for the presence of objects, this paper proposes a weakly supervised deep learning (WSDL) method for multi-class geospatial object detection using scene-level tags only. Compared to existing WSDL methods which take scenes as isolated ones and ignore the mutual cues between scene pairs when optimizing deep networks, this paper exploits both the separate scene category information and mutual cues between scene pairs to sufficiently train deep networks for pursuing the superior object detection performance. In the first stage of our training method, we leverage pair-wise scene-level similarity to learn discriminative convolutional weights by exploiting the mutual information between scene pairs. The second stage utilizes point-wise scene-level tags to learn class-specific activation weights. While considering that the testing remote sensing image generally covers a large region and may contain a large number of objects from multiple categories with large size variations, a multi-scale scene-sliding-voting strategy is developed to calculate the class-specific activation maps (CAM) based on the aforementioned weights. Finally, objects can be detected by segmenting the CAM. The deep networks are trained on a seemingly unrelated remote sensing image scene classification dataset. Additionally, the testing phase is conducted on a publicly open multi-class geospatial object detection dataset. The experimental results demonstrate that the proposed deep networks dramatically outperform the state-of-the-art methods.

1. Introduction

Multi-class geospatial object detection from remote sensing images (Cheng et al., 2014) consists of localizing objects of interest (e.g., airplanes, bridges) on the earth's surface and predicting their categories. Compared with object detection from natural images (Everingham et al., 2010; Russakovsky et al., 2015), geospatial object detection suffers from additional challenges, including large size variations, dense distributions, and arbitrary orientations (Marcos et al., 2018) of the object instances on the earth's surface. Hence, multi-class geospatial object detection requires more specific exploration.

Motivated by the great success of deep learning (Krizhevsky et al., 2012; LeCun et al., 2015), many researchers in the remote sensing community (Cheng et al., 2016; Deng et al., 2018; Ding et al., 2018; Long et al., 2017; Zhong et al., 2018; Zou and Shi, 2018) have transferred deep networks pre-trained on large-scale natural image datasets such as ImageNet (Russakovsky et al., 2015) and MSCOCO (Lin et al., 2014), to geospatial object detection. However, these geospatial object detection methods (Cheng et al., 2016; Deng et al., 2018; Ding et al., 2018; Long et al., 2017; Zhong et al., 2018; Zou and Shi, 2018) highly depend on bounding box annotations to train or fine-tune deep networks. It is well known that bounding box annotations are time-consuming and become almost impossible when the object volume is very large. As scene-level tags are much easier to collect than bounding box annotations, the past decade has witnessed major advances in constructing remote sensing image scene datasets (Cheng et al., 2017; Li et al., 2018b; Xia et al., 2017; Yang and Newsam, 2010; Zhou et al., 2018b), but the progress has been relatively slow in building geospatial object detection datasets with accurate bounding box annotations. To alleviate the labor of bounding box annotations, this paper tries to leverage the already existing remote sensing scene datasets to provide weak supervision to train deep networks for multi-class geospatial object detection.

In the existing remote sensing scene datasets (Cheng et al., 2017; Li et al., 2018b; Xia et al., 2017; Yang and Newsam, 2010; Zhou et al.,

* Corresponding author.
E-mail addresses: yansheng.li@whu.edu.cn (Y. Li), zhangyj@whu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.isprsjprs.2018.09.014
Received 19 May 2018; Received in revised form 11 September 2018; Accepted 14 September 2018

2018b), each scene contains one kind of dominant object and has varied backgrounds. Scene tags only record the category type of the dominant object in each scene, and do not contain any knowledge about the number, location, size, or orientation of the objects or backgrounds. In addition, scenes with the same tag often contain different numbers of objects with varied locations, sizes and orientations. There is no doubt that learning geospatial object detectors using the already existing remote sensing scene datasets is very cost-effective, but the learning process is very challenging because the majority of the object information is not provided.

With the aid of global pooling operations, such as global maximum pooling (GMP) and global average pooling (GAP), researchers (Zhou et al., 2014, 2016, 2018a; Oquab et al., 2015) in the computer vision community have shown that deep networks trained with only image-level/scene-level tags are informative of object locations. Unfortunately, these methods ignore the mutual information in image/scene pairs when optimizing the deep networks. In the literature, the mutual information has been widely regarded as a vital cue in the co-saliency task (Zhang et al., 2016), which also aims at collaboratively detecting common objects in multiple images. Intuitively, exploiting the mutual information in the optimization of deep networks is highly likely to improve the performance.

In this paper, we exploit the mutual information between scene pairs to train deep networks to overcome the aforementioned drawback in the existing methods (Zhou et al., 2014, 2016, 2018a; Oquab et al., 2015). With the consideration that the remote sensing image generally covers a large region and may contain many objects from multiple categories with a large size variation, we propose a multi-scale scene-sliding-voting strategy to calculate the class-specific activation maps (CAM). Furthermore, we study a set of CAM-oriented segmentation methods including a straightforward segmentation method, a diffusion-based segmentation method, and a modification-based segmentation method. As the activation maps are class-specific, it is possible to assign a suitable segmentation method for each activation map by object category, which can further improve the overall performance.

Overall, this paper trains deep networks on one large-scale remote sensing image scene classification dataset, but the learned deep networks are tested on a different multi-class geospatial object detection dataset. As can be seen, the learning supervision is extremely weak as only scene-level tags are taken as supervision and the training and testing data comes from different tasks and datasets. Even under this extreme setting, our proposed method still yields promising results, and outperforms the baselines (Oquab et al., 2015; Zhou et al., 2016). The main contributions of this paper can be summarized as follows:

- This paper proposes a new framework to train deep networks under scene-level supervision for multi-class geospatial object detection. To the best of our knowledge, this is the first method that considers the mutual information between scene pairs to train deep networks for the weak supervision scenario.
- Taking the characteristics of remote sensing images into account, we present a multi-scale scene-sliding-voting strategy to calculate the CAM of remote sensing images.
- This paper gives a set of CAM-oriented segmentation methods and analyzes their application cases, which makes selecting the best segmentation method for each activation map by object category possible.
- Last but not least, this paper reveals the use of knowledge transfer between different tasks and datasets using deep networks.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 specifically introduces how to train deep networks under scene-level supervision. Section 4 shows the multi-class geospatial object detection method using the learned deep networks under scene-level supervision. Section 5 reports the experimental results. Finally, Section 6 gives the conclusion of this paper.

2. Related work

In this section, we briefly review the most relevant works in the literature that include weakly supervised deep networks and multi-class geospatial object detection.

To alleviate the labor of bounding box annotations, pioneers in computer vision exploit scene-level or image-level tags as weak supervision for localizing objects in images or scenes. More specifically, Pinheiro and Collobert (2015) and Cinbis et al. (2017) combined multi-instance learning with deep convolutional features to localize objects. Oquab et al. (2014) proposed a method to localize objects by evaluating the output of deep networks on multiple overlapping patches. Although promising results have been reported, these methods still cannot be trained in an end-to-end way. In the most recent years, region proposals-based methods using weak supervision (Bilen and Vedaldi, 2016; Tang et al., 2017) have been proposed to address object detection. With the aid of global pooling operations, Oquab et al. (2015) and Zhou et al. (2016) trained deep networks in an end-to-end manner under weak supervision for class-specific object detection. In the most recent years, this idea has been widely explored in semantic segmentation (Chen et al., 2018; Kolesnikow and Lampert, 2016) and saliency detection (Wang et al., 2017). As these methods were originally designed for natural images, they cannot be directly used for remote sensing image analysis as they have insufficient capability to handle the challenges in remote sensing images, which contain complex backgrounds and densely distributed objects with arbitrary orientations.

In the early days, many variants of hand-crafted features have been explored to detect multi-class geospatial objects (Cheng et al., 2013, 2014; Xiao et al., 2015) under the supervision of bounding box annotations. Afterwards, many researchers (Cheng et al., 2016; Long et al., 2017; Zou and Shi, 2018) transferred deep networks pre-trained on large-scale natural image datasets to the geospatial object detection task. Although these methods achieved improved performance, they (Cheng et al., 2016; Long et al., 2017; Zou and Shi, 2018) still require bounding box annotations of geospatial objects to fine-tune the transferred deep networks. To alleviate the dependence on bounding box annotations, Han et al. (2015) proposed a probabilistic framework to jointly integrate saliency, interclass compactness, and interclass separability to initialize training instances from remote sensing images with binary labels indicating whether one image contains the objects of interest or not. In addition, the training instances were further utilized to iteratively learn object detectors. As a first effort, this approach achieved promising results on single-class geospatial object detection but could not be readily extended to the multi-class case. As reviewed in Cheng and Han (2016), how to leverage weak supervision to address multi-class geospatial object detection needs further exploration.

3. Learning deep networks under scene-level supervision

In this section, we first analyze the vulnerability of the existing weak supervision based deep networks (Oquab et al., 2015; Zhou et al., 2014, 2016, 2018a) from an architectural perspective. With the aid of a global pooling operation (e.g., GMP or GAP), existing weak supervision based deep networks adopt the architecture depicted in the second stage in Fig. 1 to learn convolutional weights and class-specific activation weights in an end-to-end manner. Due to the usage of the global pooling operation, there is only a very weak connectivity between the scene tag and convolutional layers. To facilitate understanding, we give a toy example to explain why the global pooling operation yields the weak connectivity and show the drawback of this weak connectivity in Fig. 2. In the case of the forward propagation shown in Fig. 2(a), the spatial units of each channel in the last convolutional layer are aggregated into one single unit in the aggregation feature vector. Accordingly, in the case of the backward gradient propagation shown in Fig. 2(b), the gradient value of each unit in the aggregation feature vector is equally divided into the spatial units of each corresponding