ISPRS Journal of Photogrammetry and Remote Sensing 141 (2018) 208–218

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A light and faster regional convolutional neural network for object detection in optical remote sensing images

Peng Ding[a,b,c,d,e,*], Ye Zhang[a,e], Wei-Jian Deng[b], Ping Jia[a,c], Arjan Kuijper[d]

[a] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Key Laboratory of Airborne Optical Imaging and Measurement, Chinese Academy of Sciences, Changchun 130033, China
[d] Fraunhofer Institute for Computer Graphics Research & TU Darmstadt, 64283 Darmstadt, Germany
[e] State Key Laboratory of Applied Optics, Chinese Academy of Sciences, Changchun 130033, China

ARTICLE INFO

Keywords:
Deep convolution neural network
Deep learning (DL)
Remote sensing images
Object detection

ABSTRACT

Detection of objects from satellite optical remote sensing images is very important for many commercial and governmental applications. With the development of deep convolutional neural networks (deep CNNs), the field of object detection has seen tremendous advances. Currently, objects in satellite remote sensing images can be detected using deep CNNs. In general, optical remote sensing images contain many dense and small objects, and the use of the original Faster Regional CNN framework does not yield a suitably high precision. Therefore, after careful analysis we adopt dense convoluted networks, a multi-scale representation and various combinations of improvement schemes to enhance the structure of the base VGG16-Net for improving the precision. We propose an approach to reduce the test-time (detection time) and memory requirements. To validate the effectiveness of our approach, we perform experiments using satellite remote sensing image datasets of aircraft and automobiles. The results show that the improved network structure can detect objects in satellite optical remote sensing images more accurately and efficiently.

1. Introduction

With the development of remote sensing technology, the resolution of optical remote sensing images has greatly improved and images have become largely available. Compared with other types of images, remote sensing images provide more details and a clearer texture. Thus, object detection using optical remote sensing images offers many advantages. Firstly, optical remote sensing images can be used to detect "radar stealth" objects that use surface coatings and special structures. Secondly, optical remote sensing images can provide more favorable features for detection (Cheng and Han, 2016). In the international classification competition in 2012, researchers used deep convolution neural networks (deep CNNs) to classify objects, and the precision of their approach was significantly higher than those of other methods (Guo et al., 2016). In this context, deep learning (Chen and Lin, 2014; Salakhutdinov, 2014), particularly deep CNN (LeCun et al., 2015; Schmidhuber, 2015) processing, has been applied in several fields ranging from object detection (Alshehhi et al., 2017; Fytsilis et al., 2016) to object classification (Paoletti et al., 2017; Szegedy et al., 2015; Zeiler and Fergus, 2013; Zhang et al., 2017) and tracking (Cui et al., 2016; Wang and Yeung, 2013). Different methods of reducing the network training complexity and overfitting have been presented. These include initialization from the original random distribution to those of Gauss and Xavier (Glorot and Bengio, 2010), as well as attempts to reduce the difficulty of training decline and improve convergence. Moreover, the BN (Ioffe and Szegedy, 2015) approach has been demonstrated to not only reduce training difficulty, but also the possibility of overfitting. The rectified linear unit (ReLU) and parametric ReLU (PReLU) (Glorot et al., 2011; Goodfellow et al., 2013; He et al., 2015c; Kim et al., 2015; Pan and Srikumar, 2015) activation functions have replaced the original sigmoid and tanh activation functions, and since these functions more closely resemble human biological activation, the precision of the results is greatly enhanced. In addition, the use of the dropout technique (Baldi and Sadowski, 2013; Srivastava et al., 2014) has added to the success of the deep CNN approach.

In this context we adopt in our study deep CNNs to detect objects (airplanes and automobiles) in our data sets. There are several frameworks in object detection based on deep CNNs, like Regions with CNN features (RCNN) (Girshick et al., 2014), Fast Region-based Convolutional Network (Fast RCNN) (Redmon et al., 2015), and others (Kabani and Elsakka, 2016; Sermanet et al., 2013; Zitnick and Dollar, 2014).

* Corresponding author at: Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China.
E-mail address: dingpeng14@mails.ucas.ac.cn (P. Ding).

https://doi.org/10.1016/j.isprsjprs.2018.05.005
Received 24 June 2017; Received in revised form 13 March 2018; Accepted 8 May 2018
0924-2716/ © 2018 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

Among these frameworks, the Faster RCNN approach affords suitable precision for real-time object detection (Ren et al., 2016). In the field of remote sensing, many researchers have focused on airplane detection using deep CNNs. Some of them have designed their own frameworks. Wu et al. (2015) proposed the BING approach in combination with a CNN to perform aircraft detection. However, the average detection time (or test time) for test images with this approach is about 6.414 s. In addition, the precision is not that high. Along similar lines, Cao et al. (2016) performed airplane detection by means of RCNN, which is thought to perform poorer than Faster RCNN in terms of both precision and speed. Zhang et al. (2016) performed aircraft detection by using weakly supervised CNNs. This approach is similar to the RPN + Fast RCNN (Faster RCNN without feature sharing) approach.

In the field of remote sensing, many researchers have also performed vehicle detection using deep CNNs. Ammour et al. performed car detection by combining CNNs and support vector machines (SVMs), similar to the RCNN approach (Ammour et al., 2017). Tang et al. performed vehicle detection by using RCNNs and Hard Negative Example Mining (Tang et al., 2017), which is an improvement on the Faster RCNN. They performed vehicle detection by adapting ZF-Net as the baseline and using the RealBoost algorithm to replace the Fast RCNN.

Our work is different from these approaches, since we adopt a more advanced framework, the Faster RCNN (Ren et al., 2016) framework, and choose the VGG16 network (Simonyan and Zisserman, 2015), a very deep CNN network, as the base network to detect objects. So Faster RCNN forms the holistic framework and VGG16-Net is the base network used in this framework. To improve the precision and recall of the tests, we adopt specific measures to strengthen the capability of VGG16-Net. Since the computational cost is a major problem that restricts Faster RCNN applications, we propose the use of a fully convolutional neural network instead of the fully connected layers in the Faster RCNN framework. Through this approach, the memory requirements of the final model become significantly smaller. The test-time also reduces considerably. Moreover, the precision of the approach is still able to meet our requirements.

The main contributions of this paper are thus as follows:

1. 1 For the detection of dense objects in optical remote sensing images, we adopt dilated convolutions instead of traditional convolutions to improve precision.
2. As certain objects in satellite remote sensing images are small and difficult to detect, we adopt a bootstrapping strategy called Online Hard Example Mining (Shrivastava et al., 2016) for mining hard negative examples, and we add it to Faster RCNN.
3. We use a multi-scale representation and its combinations in a new manner.
4. We propose a fully convolutional neural network instead of the fully connected layers in the Faster RCNN framework.
5. The object detection accuracy and recall show significant improvement with our approach.

The rest of the paper is organized as follows: In the next section, we describe the basic principles of CNNs and the development and principles of Faster RCNN. The details of our method are explained in Section 3. Our analysis and comparison of experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

2.1. Principles of convolutional neural networks

Traditional CNNs are composed of multiple stages, with each stage consisting of a convolution layer, a feature pooling layer, and a fully connected (FC) layer (Krogh and Hertz, 1992; Lecun et al., 1998).

Convolution layers: At the convolution layer, the previous layer's feature maps X_i^{l-1} are convolved with learnable kernelsk_{ij}^l, a trainable bias parameterb_j is added and the result is processed by the activation function $f(\cdot)$ to form the output feature map. This process can be expressed as:

$$X_j^l = f\left(\sum_{i \in Mj} X_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

Here, M_j represents a selection of input maps. In this work, we chose ReLU, which is called the rectifier activation function, as the activation function in the new layers since it works better than the logistic sigmoid and hyperbolic tangent functions (Glorot et al., 2011).

Feature pooling layer: This layer treats each feature map separately. In general, this layer is called the subsampling layer, and it produces down-sampled versions of the input maps. This means that the number of input and output maps is the same, but the output maps are smaller in size. The results are robust to small variations in the location of features in the previous layer. This process can be expressed as:

$$X_j^l = down(X_j^{l-1}) \tag{2}$$

Here, $down(\cdot)$ denotes a down-sampling operation. By means of down-sampling, we reduce the size of the input by summarizing neurons from a small spatial neighborhood (Scherer et al., 2010).

Fully connected (FC) layers: After data processing by several convolutional and subsampling layers, high-level reasoning in the neural network is performed via FC layers. Neurons in an FC layer have full connections to all activations in the previous layer. Their activations can hence be computed with a matrix multiplication followed by a bias offset. The flowchart of a CNN is shown in Fig. 1.

Training is performed by means of the backpropagation algorithm (Chen et al., 2008) to minimize the aberrations between the ideal output and the actual output of the CNNs. In general, for the purpose of detection, a CNN is followed by a classification module.

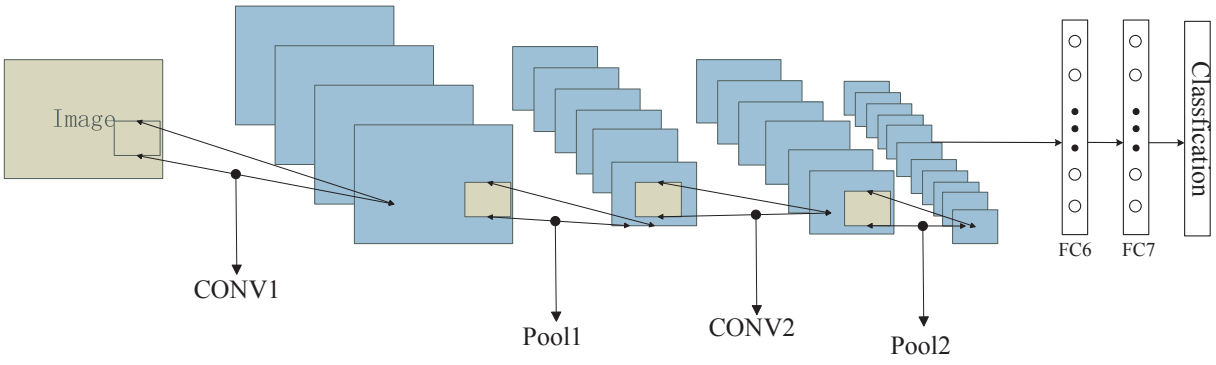

Fig. 1. Flowchart of convolutional neural network.

Download English Version:

https://daneshyari.com/en/article/6949133

Download Persian Version:

https://daneshyari.com/article/6949133

Daneshyari.com