ISPRS Journal of Photogrammetry and Remote Sensing 145 (2018) 60–77

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning

Ronald Kemker, Carl Salvaggio, Christopher Kanan*

Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, 54 Lomb Memorial Drive, Rochester, NY 14623, United States

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural network
Semantic segmentation
Multispectral
Unmanned aerial system
Synthetic imagery

ABSTRACT

Deep convolutional neural networks (DCNNs) have been used to achieve state-of-the-art performance on many computer vision tasks (e.g., object recognition, object detection, semantic segmentation) thanks to a large repository of annotated image data. Large labeled datasets for other sensor modalities, e.g., multispectral imagery (MSI), are not available due to the large cost and manpower required. In this paper, we adapt state-of-the-art DCNN frameworks in computer vision for semantic segmentation for MSI imagery. To overcome label scarcity for MSI data, we substitute real MSI for generated synthetic MSI in order to initialize a DCNN framework. We evaluate our network initialization scheme on the new RIT-18 dataset that we present in this paper. This dataset contains very-high resolution MSI collected by an unmanned aircraft system. The models initialized with synthetic imagery were less prone to over-fitting and provide a state-of-the-art baseline for future work.

1. Introduction

Semantic segmentation algorithms assign a label to every pixel in an image. In remote sensing, semantic segmentation is often referred to as image classification, and semantic segmentation of non-RGB imagery has numerous applications, such as land-cover classification (IEEE GRSS, 2017), vegetation classification (Laliberte et al., 2011), and urban planning (Rottensteiner et al., 2012; Volpi and Ferrari, 2015). Semantic segmentation has been heavily studied in both remote sensing and computer vision. In recent years, the performance of semantic segmentation algorithms for RGB scenes has rapidly increased due to deep convolutional neural networks (DCNNs). To use DCNNs for semantic segmentation, they are typically first trained on large image classification datasets that have over one million labeled training images. Then, these pre-trained networks are then adapted to the semantic segmentation task. This two-step procedure is necessary because DCNNs that process high-resolution color (RGB) images have millions of parameters, e.g., VGG-16 has 138 million parameters (Simonyan and Zisserman, 2014). Semantic segmentation datasets in computer vision are too small to find good settings for the *randomly initialized* DCNN parameters (weights), and over-fitting would likely occur without the use of pre-trained networks. For example, to evaluate a semantic segmentation method on the RGB PASCAL VOC datasets (Everingham et al., 2010), state-of-the-art methods use a DCNN pre-trained on ImageNet (1.28 million training images), fine-tune it for semantic segmentation on the COCO dataset (80 K training images) (Lin et al., 2014), and then fine-tune it again on PASCAL VOC (1464 training images) (Chen et al., 2018; Lin et al., 2017).

Utilizing pre-trained networks to prevent overfitting works well for RGB imagery because massive labeled datasets are available; but in the non-RGB domain, label scarcity is a far greater problem. For example, existing semantic segmentation benchmarks for hyperspectral imagery consist of a single image mosaic. Therefore, pre-training DCNNs on hand-labeled datasets consisting of real images is not currently possible in non-RGB domains. In this paper, we explore an alternative approach: using vast quantities of automatically-labeled *synthetic* multispectral imagery (MSI) for pre-training DCNN-based systems for semantic segmentation.

We propose to use the Digital Imaging and Remote Sensing Image Generation (DIRSIG) modeling software to generate large quantities of synthetic MSI and corresponding label maps. We use DIRSIG to build a large, diverse scene model, in which we can simulate various weather and lighting conditions. We then capture synthetic aerial images of the scene with a MSI sensor model. We use the synthetic data to initialize a DCNN for object recognition, and then we combine the pre-trained DCNN with two different fully-convolutional semantic segmentation models using real MSI (see Fig. 1).

In the past, researchers have used DCNNs pre-trained on ImageNet to yield state-of-the-art results for the semantic segmentation of high-resolution multispectral aerial imagery (Wang et al., 2017; Marmanis et al., 2016) because the most widely used benchmarks (Rottensteiner et al., 2012) only use a single non-RGB band. What happens when the

* Corresponding author.
E-mail addresses: rmk6217@rit.edu (R. Kemker), cnspci@rit.edu (C. Salvaggio), kanan@rit.edu (C. Kanan).

https://doi.org/10.1016/j.isprsjprs.2018.04.014
Received 10 June 2017; Received in revised form 6 March 2018; Accepted 20 April 2018
Available online 30 April 2018

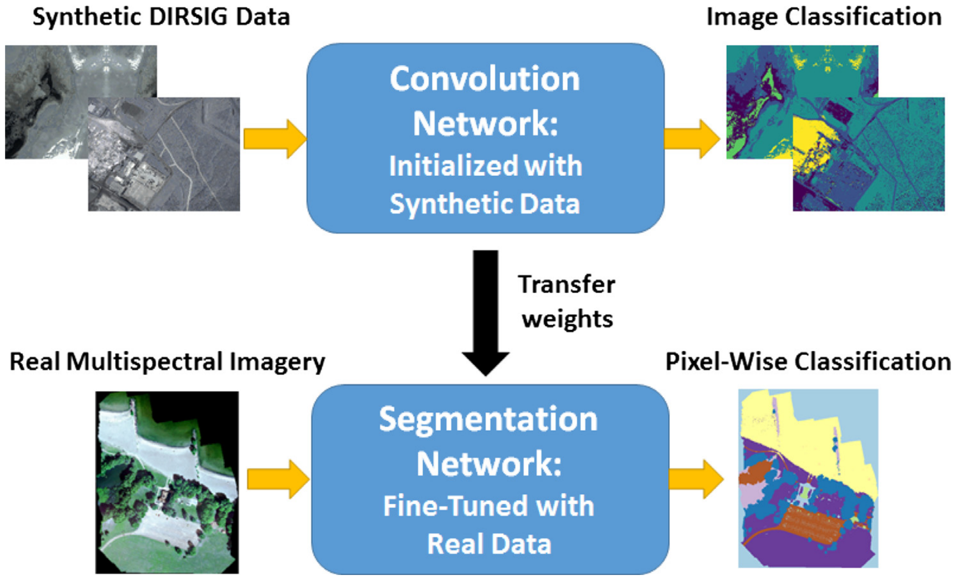

Fig. 1. Our proposed model uses synthetic multispectral imagery to initialize a DCNN for semantic segmentation. This model is then fine-tuned on real imagery.

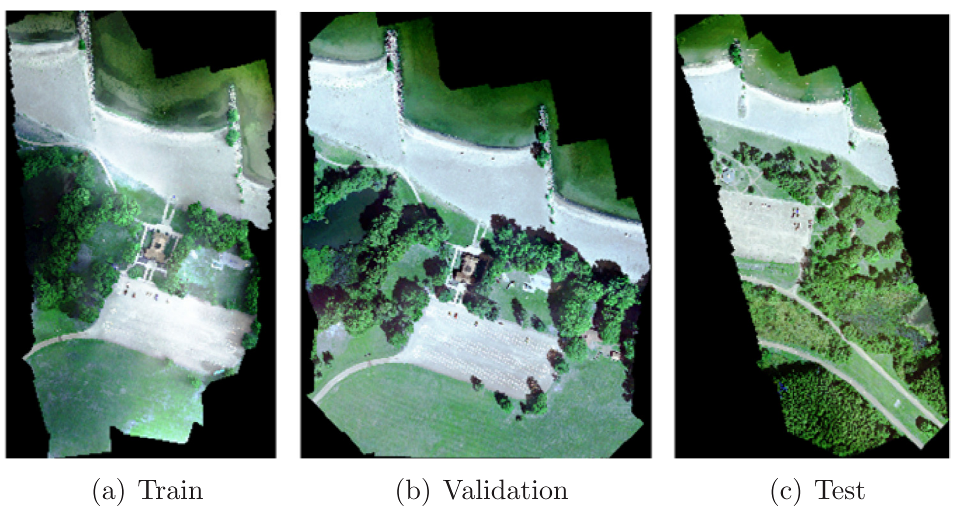

Fig. 2. RGB visualization of RIT-18 dataset. This dataset has six spectral bands.

spectral range of the dataset increases? The real MSI used to evaluate our network initialization scheme comes from a new semantic segmentation dataset that we built called RIT-18.[1] RIT-18 consists of high-resolution MSI (six bands) acquired by an unmanned aircraft system (UAS). The primary use of this dataset is for evaluating semantic segmentation frameworks designed for non-RGB remote sensing imagery. The dataset, shown in Fig. 2, is split into training, validation, and testing folds to (1) provide a standard for state-of-the-art comparison, and (2) demonstrate the feasibility of deploying algorithms in a more realistic setting. Baseline results demonstrate that the large spatial variability commonly associated with high-resolution imagery, large sample (pixel) size, small and hidden objects, and unbalanced class distribution make this a difficult dataset to perform well on, making it an excellent dataset for evaluating our DCNN frameworks for semantic segmentation.

Contributions: Our paper makes three major contributions: (1) We are the first to adapt recent fully-convolutional DCNNs to semantic segmentation of multispectral remote sensing imagery; (2) We demonstrate that pre-training these networks on synthetic imagery can significantly improve their performance; and (3) We describe the new RIT-18 dataset for evaluating MSI semantic segmentation algorithms.

2. Related work

2.1. Semantic segmentation of RGB imagery with deep networks

In this paper, pixel-wise classification and semantic segmentation are synonymous. Semantic segmentation is the term more commonly used in computer vision and is becoming increasingly used in remote sensing. State-of-the-art semantic segmentation frameworks for RGB imagery are trained end-to-end and consist of convolution and segmentation sub-networks. The convolution network is usually a pre-trained DCNN designed to classify images from ImageNet (Long et al., 2015; Noh et al., 2015; Chen et al., 2018; Pinheiro et al., 2016), and current state-of-the-art performers use VGG-16 (Simonyan and Zisserman, 2014) or ResNet (Russakovsky et al., 2015). The segmentation network is appended to the convolution network and is designed to reconstruct the feature response to the same spatial dimensions as the input before assigning semantic labels. The resulting semantic segmentation network can be fine-tuned with orders of magnitude fewer training images (thousands versus millions of images) because

[1] The dataset is available at https://github.com/rmkemker/RIT-18.