ISPRS Journal of Photogrammetry and Remote Sensing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models

Diego Marcos [a,b,*], Michele Volpi [a], Benjamin Kellenberger [b], Devis Tuia [a,b]

[a] MultiModal Remote Sensing, University of Zurich, Switzerland[1]
[b] Laboratory of GeoInformation Science and Remote Sensing, Wageningen University and Research, The Netherlands[2]

ARTICLE INFO

Article history:
Received 11 June 2017
Received in revised form 25 November 2017
Accepted 27 January 2018
Available online xxxx

Keywords:
Semantic labeling
Deep learning
Rotation invariance
Sub-decimeter resolution

ABSTRACT

In remote sensing images, the absolute orientation of objects is arbitrary. Depending on an object's orientation and on a sensor's flight path, objects of the same semantic class can be observed in different orientations in the same image. Equivariance to rotation, in this context understood as responding with a rotated semantic label map when subject to a rotation of the input image, is therefore a very desirable feature, in particular for high capacity models, such as Convolutional Neural Networks (CNNs). If rotation equivariance is encoded in the network, the model is confronted with a simpler task and does not need to learn specific (and redundant) weights to address rotated versions of the same object class. In this work we propose a CNN architecture called Rotation Equivariant Vector Field Network (RotEqNet) to encode rotation equivariance in the network itself. By using rotating convolutions as building blocks and passing only the values corresponding to the maximally activating orientation throughout the network in the form of orientation encoding vector fields, RotEqNet treats rotated versions of the same object with the same filter bank and therefore achieves state-of-the-art performances even when using very small architectures trained from scratch. We test RotEqNet in two challenging sub-decimeter resolution semantic labeling problems, and show that we can perform better than a standard CNN while requiring one order of magnitude less parameters.

1. Introduction

In this paper we consider the task of *semantic labeling*, which corresponds to the automatic assignment of each pixel to a set of predefined land-cover or land-use classes. The classes are selected specifically for the task to be solved and define the learning problem for the model.

When using low- to mid-resolution multispectral imagery (e.g. Landsat), it is customary to assume that the spectral information carried by a pixel is sufficient to classify it into one of the semantic classes, thus reducing the need for modeling spatial dependencies. However, when dealing with very-high spatial resolution (VHR) imagery, i.e. imagery in the meter to sub-decimeter resolution range, the sensor trades off spectral resolution to gain spatial details. Such data is commonly composed of red-green-blue (RGB) color channels, occasionally with an extra near infrared (NIR) band. Due to this trade-off, single pixels tend not to contain sufficient information to be assigned with high confidence to the correct semantic class, when relying on spectral characteristics only. Moreover, depending on the task, some classes can be semantically ambiguous: a typical example is land use mapping, where objects belonging to different classes can be composed of the same material (e.g. road and parking lots), thus making analysis based on spectra of single pixels not suitable. To resolve both problems, spatial context needs to be taken into account, for example via the extraction and use of textural (Regniers et al., 2016), morphological (Dalla Mura et al., 2010; Tuia et al., 2015), tree-based (Gueguen and Hamid, 2015) or other types (Malek et al., 2014) of spatial features. These features consider the neighborhood around a pixel as part of its own characteristics, and allow to place spectral signatures in context and solve ambiguities at the pixel level (Fauvel et al., 2013). The diverse and extensive pool of possible features led to a surge in works focusing on the automatic generation and selection of discriminant features (Harvey et al., 2002; Glocer et al., 2005; Tuia et al., 2015), aimed at preventing to compute and store features that are redundant or not suited for a particular task.

[1] www.geo.uzh.ch/en/units/multimodal-remote-sensing.
[2] www.geo-informatie.nl.

* Corresponding author at: Laboratory of Geo-information Science and Remote Sensing, PO Box 47 6700 AA Wageningen, Netherlands
E-mail address: diego.marcos@wur.nl (D. Marcos).

https://doi.org/10.1016/j.isprsjprs.2018.01.021
0924-2716/

Another common approach to reduce the computational burden while enforcing spatial reasoning is to extract local features from a support defined by unsupervised segmentation. Also, spatial rules can be encoded by Markov random fields, where spatial consistency is usually enforced by minimizing a neighborhood-aware energy function (Moser et al., 2013) or specific spatial relationships between the classes (Volpi and Ferrari, 2015).

In the situations described above, a successful solution comes at the cost of having to manually engineer a high-dimensional set of features potentially covering all the local variations of the data in order to encode robust and discriminative information. In this setting, there is no guarantee that the features employed are optimal for a given semantic labeling problem. These problems raised the interest of the community in solutions avoiding to manually engineer the feature space, solutions that are extensively studied under the *deep learning* paradigm. The aim of deep learning is to train a parametric system learning feature extraction *jointly* with a classifier (Goodfellow et al., 2016), in an end-to-end manner. When focusing on image data, Convolutional Neural Networks (CNNs, LeCun et al. (1998)) are state-of-the-art. Their recognized success follows from new ground-breaking results in many computer vision problems. CNNs stand out thanks to their ability to learn complex problem-specific features, while jointly optimizing a loss (e.g. a classifier, a regressor, etc.). Thanks to recent hardware advances accelerating CNN training consistently, as well as the existence of pre-trained models to get started, CNNs have become one of the most studied models in recent remote sensing research dealing with VHR imagery, as we briefly review below.

The first models proposed studied the effectiveness of translating computer vision architectures directly to aerial data for tile classification. In that sense, a single label was retrieved per image tile, thus tackling what in computer vision is called the *image classification problem*[3]: authors in Castelluccio et al. (2015) and Penatti et al. (2015) studied the effect of fine-tuning models trained on natural image classification problems, in order to adapt them quickly to above-head image classification. Their results suggested that such a strategy is relevant for image classification and can be used to reuse models trained on a different modality. Transposing these model in the semantic labeling problem is also possible, typically applying the models using a sliding window centered at each location of the image, as tested in Campos-Taberner et al. (2016). However, the authors also came to three important conclusions: (i) models trained from scratch (in opposition to fine-tuned models from vision) tend to provide better results on specific labeling tasks; (ii) by predicting a single label per patch, the one corresponding to the pixel on which the patch is centered, these models are not able to encode explicit label dependencies in the output space and (iii) the computational overhead of the sliding window approach is extremely large. Such conclusions support the use of network architectures that have been developed specifically for semantic labeling problems: recent efforts tend to consider *fully convolutional* approaches (Long et al., 2015), where the CNN does not only predict a single label per patch, but actually provides directly the label map for all the pixels that compose the input tile. The approaches proposed vary from spatial interpolation (Maggiori et al., 2017), fully convolutional models (Audebert et al., 2016), deconvolutions (Volpi and Tuia, 2017), stacking activations (Maggiori et al., 2016) to hybridization with other classifiers (Liu et al., 2017), but they all are consistent in one observation: fully convolutional architectures drastically reduce the inference time and naturally encode some aspect of output dependencies, in particular learning dependent filters at different scales, thus reducing the need of cumbersome postprocessing of the prediction map.

While these works open endless opportunities for remote sensing image processing with CNNs, they also showed one of the biggest downsides of these models: CNNs tend to need large amounts of ground truth to be trained, and setting up the architecture, as well as selecting hyperparameters, can be troublesome, since cross-validation is often prohibitive in terms of processing time. Note that it is often that case when the number of parameters is larger than the number of training samples, which makes regularization techniques and data augmentation a must-do, at the cost of significantly slowing model training. Our contribution aims at addressing this drawback of CNNs, *i.e.* the large model sizes and need for labels when there is a limited availability of ground truth. In this paper, we propose to tackle the problem by exploiting a property of objects and features in remote sensing images: *their orientation is arbitrary*.

Overhead imagery differs from natural images in that the *absolute* orientation of objects and features within the images tends to be irrelevant for most tasks, including semantic labeling. This is because the orientation of the camera in nadir-looking imagery is most often arbitrary. As a consequence, the label assigned to an element in the image should not change if the image is taken with a different camera orientation. We call this property *equivariance*, and it is a property that recently attracted a lot of interest in image analysis (Lei et al., 2012; Cheng et al., 2016).

Given a rotation operator, $g_\alpha(\cdot)$, we say that a function $f(\cdot)$ is equivariant to rotations if $f(g_\alpha(\cdot)) = g_\alpha(f(\cdot))$, invariant to rotations if $f(g_\alpha(\cdot)) = f(\cdot)$ and, more generally, covariant to rotations if $f(g_\alpha(\cdot)) = h(f(\cdot))$, with $h(\cdot)$ being some function other than $g_\alpha(\cdot)$. Note that, in the case of semantic labeling, the property we are interested in is equivariance, although it becomes invariance if we consider a single pixel at a time. We will therefore use the terms equivariance and invariance interchangeably in this paper.

With CNNs, equivariance to the rotation of inputs can be approximated by randomly rotating the input images during training, a technique known as *data augmentation* or *jittering* (Leen, 1995). If the CNN has enough capacity and has seen the training samples in sufficient number of orientations, it will learn to be invariant to rotations (Lenc and Vedaldi, 2015). While this kind of data augmentation greatly increases the generalization accuracy, it does not offer any advantage in terms of model compactness, since similar filters, but with different orientations, need to be learned independently. A different approach, hard coding such invariances within the model, has the two main beneficial effects: first, the model becomes robust to variations which are not discriminative, as a standard CNN with enough filters would learn; and second, model-based invariance can be interpreted as some form of regularization (Leen, 1995). This added robustness ultimately lead to models which have high capacity (as high as a standard CNN) but with lower sample complexity.

There has been a recent surge in works that explore ways of encoding model-based rotation invariance in CNNs. Laptev et al. (2016) perform a rotation of the input image in order to reduce the sample complexity of the problem and Jaderberg et al. (2015) extend this to affine transformations. These approaches provide invariance to a global rotation of the input image and not to local relative rotations, and are therefore not very well suited for segmentation tasks. Cohen and Welling (2016) encode equivariance to shifts and to rotations by multiples of 90° by tying filter weights, while Zhou et al. (2017) use linearly interpolated filters. These two methods are in principle suited for segmentation tasks. The former is limited to invariance to 90° rotations and the latter, although offering more flexibility, has the drawback of requiring a trade-off between the number of rotations and the memory requirements, bringing the authors to use 8 orientations, at multiples of 45°. Worrall et al. (2016) reduce the space of possible filters to

[3] This is not to be confused with the *semantic labeling* problem we address in this paper, which is the task of attributing a label to every pixel in the tile.

Please cite this article in press as: Marcos, D., et al. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.021