DOMAIN ADAPTATION FOR LARGE SCALE CLASSIFICATION OF VERY HIGH RESOLUTION SATELLITE IMAGES WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

T. Postadjian\textsuperscript{a}, A. Le Bris\textsuperscript{a}, H. Sahbi\textsuperscript{b}, C. Mallet\textsuperscript{a}

\textsuperscript{a}Univ. Paris Est, LASTIG MATIS, IGN, ENSG, F-94160 Saint-Mande, France
\textsuperscript{b}CNRS, LIP6 UPMC Sorbonne Universités, Paris, France

ABSTRACT

Semantic segmentation of remote sensing images enables in particular land-cover map generation for a given set of classes. Very recent literature has shown the superior performance of deep convolutional neural networks (DCNN) for many tasks, from object recognition to semantic labelling, including the classification of Very High Resolution (VHR) satellite images. However, while plethora of works aim at improving object delineation on geographically restricted areas, few tend to solve this classification task at very large scales. New issues occur such as intra-class class variability, diachrony between surveys, and the appearance of new classes in a specific area, that do not exist in the predefined set of labels. Therefore, this work intends to (i) perform large scale classification and to (ii) expand a set of land-cover classes, using the off-the-shelf model learnt in a specific area of interest and adapting it to unseen areas.

Index Terms— Land-cover mapping, satellite images, Very High Spatial Resolution, large-scale, learning, domain adaptation, deep neural networks, land-cover geodatabases.

<table>
<thead>
<tr>
<th>conv1</th>
<th>MP</th>
<th>conv2</th>
<th>MP</th>
<th>conv3</th>
<th>MP</th>
<th>conv4</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>2x2</td>
<td>32</td>
<td>2x2</td>
<td>64</td>
<td>2x2</td>
<td>128</td>
<td>2x2</td>
</tr>
</tbody>
</table>

Table 1: The architecture of the neural network.

1. INTRODUCTION

The automatic retrieval of a class for each pixel of Very High Resolution (VHR) geospatial images is a problem that the state-of-the-art approaches usually deal with through machine learning based supervised methods \cite{1}. Such approaches are fed with pre-computed hand-crafted features that often require expert knowledge (e.g. spectral indices such as NDVI for vegetation detection). Features are designed to capture and model the characteristics of each class during a training step. Hence, such features have to be discriminative to separate the various classes of interest. In addition, the training set must be designed carefully so as to reflect each intra-class variability. Both aspects are enhanced when classifying mono-temporal VHR images with superior textural content and intra-class variabilities: feature extraction and training set design have a crucial impact on the classification performance. At large scales, this become even more complicated as classes exhibit significant spectral fluctuations and some classes tend to show similar characteristics. In addition, a geographic large scale may yield changes in the set of classes to consider (from urban to rural to mountains for instance).

The advantage of Deep Convolutional Neural Networks (DCNNs) mainly lies in the fact that they generalize information well. This comes at a high cost in training data, so as to pinpoint each class characteristics and instances. If standard classifiers of the past decades need hand-crafted features, DCNNs extract its own features from the training dataset, in an end-to-end manner during the training step. This makes them very robust and suitable for large scale purposes: domain adaptation \cite{2, 3} can be conducted by simply providing new training samples which allows to fine-tune the computed architecture to a specific area with specific spatial and semantic characteristics.

DCNNs have been widely used in overhead imagery scene parsing, often dealing with multi-modal data such as a Digital Surface Models (DSM), lidar point clouds, heterogeneous datasets (topographic databases), or multiple deep-based architectures.

Current approaches are all based on the Fully Convolutional Networks (FCN), in opposition to patch-based approaches, first introduced by \cite{4}: \cite{5} fed both DSM and aerial imagery to a customized FCN denoted as no-downsampling FCN due to the absence of the so-called max-pooling layers. This preserves the input size. The work of \cite{6} pointed out different strategies as for the use of DCNNs and compared them in terms of performance and computation times, preferring FCN strategy over patch-based strategy. On the other hand, efforts have been made for joint edge detection and semantic classification \cite{7}, for better object delineation.

This work focuses more on a supervised domain adaptation problem, whether it is temporal, modal or in terms of semantics. The overall perspective is labeling very large regions in reasonable times. We aim to evaluate how easy it is to extent the model over wide areas with potentially new classes and potentially asynchronous scenes.

2. DOMAIN ADAPTATION

Seeking large scale classification means dealing with different sets of classes, regarding the geographic region or the goal aimed by the classification. However, training as many new net models as regions, dates, images or classification problems would be inefficient. As mentioned in Section 1, domain adaptation in deep learning consists in re-using an existing net over new areas, or with different classes, bringing slight modifications that can be apprehended by fine-tuning this net. For instance, it requires a base model that was trained over region A and a limited amount of new training samples over...
3. DATA

3.1. Satellite images

Images acquired by the SPOT 6 and 7 satellite sensors are used. They contain 4 spectral bands: Red, Green, Blue, Infra-Red at the spatial resolution of 1.5 m, once pan-sharpened. They are provided freely through the French Theia land data service. Each ground pixel is covered by a single image every year, offering a mono-temporal very high spatial resolution analysis. Such study is complementary to high resolution multi-temporal large-scale land-cover analysis, that became possible with the Landsat and Sentinel programs [9]. Moreover, contrary to aforementioned works, no Digital Surface Model is inserted so as to discriminate roads and buildings as it comes at high cost and cannot be reasonably considered in a large scale perspective. Moreover, our previous works shown that it is not necessary in order to obtain satisfactory discrimination performance.

3.2. Land-cover geodatabases

The possibility to use DCNNs for large-scale land-cover mapping arose jointly with the national and freely available geo-databases at our disposal, providing a huge source of training data. They consist in indexing land-cover objects with their geographic location. Although, such databases are subject to errors, be it mislabeling or geometric errors (due to slightly outdated and generalized data). Such errors do not occur so frequently when considering the amount of information involved during training.

3.3. DCNN architecture

A simple yet effective architecture is used. It is provided in Table 1. Input images are of size of $65 \times 65 \times 4$ (number of bands). Only convolutions of size $3 \times 3$ are used to limit the number of parameters. The second line of the table displays the number of filters per convolution layer. The ReLU activation function is set after each convolution in order to introduce non-linearity and max-pooling layers increase the receptive field (the spatial information taken into account by the filter). Finally, a fully-connected layer on top sums up the information contained in all features in the last convolution layer.

4. EXPERIMENTS

4.1. Geographic and temporal adaptation

The domain adaptation was first tested with a net, defined as in Section 3.3, that was trained from scratch over French Brittany images acquired in 2014, using 10,000 samples for each of the following five classes: buildings, vegetation, roads, crops, water. This net was then applied in a straightforward way to Gironde images acquired in 2016, where those classes have strong variability and different behaviors compared to their representation in Brittany, and to 2014 (different time of acquisition during the year yielding different growth profile for instance for the vegetation class). Fine-tuning was also applied to assess this approach of domain adaptation. Figure 2 shows the region we focused on for this test, while Figure 3 and Figure 4 respectively show the results of direct application of a pre-trained network on a different area, and the result of the fine-tuning approach, along with Kappa index represented by heat maps over the region. Each pixel is the kappa computed over a $2,000 \times 2,000$ sub-region of the initial region. Using an Nvidia GTX 1080 Ti GPU, training from scratch over French Brittany took one day, while fine-tuning needed only 2 hours with 3,000 new samples instead of 10,000 samples per class. Fine-tuning makes overall substantial improvements visually and quantitatively: the estuary region B that is to be classified. Further experiments have shown that not only it requires few training samples compared to a "from scratch" training, but the training also converges faster [8]. Additionally, one is often confronted to the task of mapping an area that contains some objects of a class that is wanted to be retrieved, but with a poor representativeness of this class on this area. Training the net with instances of this class from another region with a fine-tuning with the few available samples can be led to solve this problem. For instance, it is the case with the task of detecting buildings in rural regions where such objects are very rare.
Fig. 2: SPOT image used for the experiment on the geographic and temporal adaptation.

<table>
<thead>
<tr>
<th>building</th>
<th>road</th>
<th>vegetation</th>
<th>crops</th>
<th>water</th>
<th>hedge</th>
</tr>
</thead>
<tbody>
<tr>
<td>73.30</td>
<td>80.85</td>
<td>83.66</td>
<td>95.84</td>
<td>67.76</td>
<td>x</td>
</tr>
<tr>
<td>72.15</td>
<td>79.94</td>
<td>84.28</td>
<td>92.36</td>
<td>68.98</td>
<td>43.54</td>
</tr>
</tbody>
</table>

Table 2: F-scores evolution with the additional hedge class.

4.2. Nomenclature shift

Expanding the number of classes is a common expectation from users of land-cover maps. However, this is rarely addressed in an effective way: one would re-construct from scratch the training set design as well as the training itself to provide new maps. Adapting the nomenclature seems far more efficient. In practice, such fine-tuning was performed to get better results on the road class since it was often confused with the vegetation. Indeed, analyzing these mislabeling errors showed that the vegetation it got confused with was mainly hedges. Therefore, a class hedge was added to cope with these ambiguities (Figure 1). Table 2 shows that we could alleviate the vegetation-roads ambiguities since the road class gets a better F-score. It can be noted that other classes, apart water, seem to decrease in terms of quantitative performance, while the visual aspect does not vary. In particular, the crop loses 3.5 points: this is due to the ground truth for the hedge, which is the only one we have, but is not up-to-date, therefore considers actual hedges as crops. But a visual inspection reveals that the classification correctly detects new hedges where there was previously crops. The road and hedge classes also have confusions: because hedges often mark roads out, the borders between them is fuzzy with such a resolution. In addition, roads are thin objects, so misclassification errors on their borders have a strong impact on the final F-score. Finally, the 5-class labeling already had a poor urban vegetation description, and this ambiguity affects the vegetation-building and hedge-building discriminations in the 6-class labeling. The main objective which was to retrieve roads more accurately is achieved with overall satisfying classification, though most likely better than numbers seem to show since the reference data is not up-to-date.

5. CONCLUSION

DCNNs are a very efficient classification tool, but training a net can be long and training data greedy. However, DCNN models are perfectly adapted for domain adaptation: fine-tuned them enables to train faster a good model from a limited training sets. Experiments have shown such approach could be performed successfully to cope with both geographic and temporal adaptation (other landscape, other season), as well as semantic adaptation (different legend), enabling an operational and efficient process of VHR satellite images for large scale classification.

6. ACKNOWLEDGEMENTS

This work was supported by public funds received in the framework of GEOSUD, a project (ANR-10-EQPX-20) of the program “Investissements d’Avenir” managed by the French National Research Agency.

7. REFERENCES


(a) Before geographic and temporal adaptation:
- buildings
- roads
- crops
- forest
- water

(b) Kappa index:
the overall result is pretty inaccurate with strong confusions building-water and building-vegetation.

**Fig. 3:** Geographic and temporal shift by direct application of a pre-trained net

(a) After domain adaptation:
- buildings
- roads
- crops
- forest
- water

(b) Kappa index:
overall substantial improvements; the urban description is poor, revealing that fine-tuning might be reinforced in this area.

**Fig. 4:** Geographic and temporal shift after fine-tuning.