Effective Building Extraction by Learning to Detect and Correct Erroneous Labels in Segmentation Mask
Praveer Singh, Nikos Komodakis

To cite this version:
Praveer Singh, Nikos Komodakis. Effective Building Extraction by Learning to Detect and Correct Erroneous Labels in Segmentation Mask. IGARSS, 2018, Valencia, Spain. hal-01832798

HAL Id: hal-01832798
https://hal-enpc.archives-ouvertes.fr/hal-01832798
Submitted on 11 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
Semantic segmentation is pivotal for remote sensing image analysis. Although existing segmentation techniques perform well on similar landscape images, their generalization capability on an entirely different landscape is extremely poor. One of the primary reasons is that they partially or wholly, neglect the underlying relationship that exist in the joint space of input and output variables. Thus, effectively they lack to impose structure in their output predictions which is necessary for successful segmentation. In this paper, we address this problem and propose a novel solution by modeling the joint distribution of input-output variable which in turn enforces some structure in the initial segmentation mask. To this end, we first detect erroneous labels, in the form of Error maps, in the initial building masks. These Error maps are then used to correct the corresponding erroneous labels through a replacement technique. We evaluate our methodology on the benchmark Inria Aerial Image Labeling dataset, which is a large scale high resolution dataset for building footprint segmentation. In contrast to previous methods, our predicted segmentation masks are much closer to ground truth, owing to the fact that they are able to effectively correct both the large errors as well as the blobby effects. We lastly perform on par with other state-of-the-arts, validating the efficacy of our technique.

Index Terms— Deep learning, high-resolution imagery, semantic segmentation, structured prediction, building footprint extraction.

1. INTRODUCTION
Recently, we have witnessed an explosion of petabytes of high-resolution remote sensing datasets such as Sentinel 1-5 [2], SpaceNet [3] and Inria Aerial Image Labeling dataset [1]. Till now, these datasets have been manually annotated by experts. However, with the availability of such an enormous amount of high-resolution data, it is truly herculean to label all of them by hand. This necessitates the automatic segmentation of these remote sensing images to quickly and effectively detect varied points of interest such as roads, buildings, forests in an image for tasks such as urban scene planning, green cover monitory and other emergency relief operations such as floods, forest fires and cyclones. Deep learning methods have recently shown significant improvements on automatic semantic labeling tasks for classical datasets such as Vaihingen or Potsdam. [4, 5] fused semantic maps from multiple sources through a residual correction technique. Whereas, [6] corrected shifts in OSM maps through an iterative refinement technique using Recurrent Neural networks. In addition, [7] used boundary detections to improve the semantic segmentation and report impressive performance on Vaihingen dataset.

Interestingly, [1] showcased the drawback of models trained over such classical datasets, by highlighting their lack of generalization capability to other cities that are captured under different conditions. They, henceforth build a new dataset covering a much larger surface of the earth including both densely populated urban landscape as well as sparse alpine regions under varied illumination conditions. Subsequently, they split their train and test data such that they come from different cities. Finally, they also addressed problem of blobby-predictions i.e., curvy edges of building masks. By fusing feature maps from different levels of convolution network, both [1] and [8] combine low level edge information with high level semantic rich information. For the same task, [9] uses a multi-task loss in order to approximate the distance transform and the semantic maps.

Lately, most of the aforementioned techniques have partially or fully neglected the idea of enforcing structure in the output label space. This mainly results in these blobby ef-
effects due to naive up-sampling during the segmentation masks prediction. On the contrary, a few of those that do enforce structure, either do it through a novice human assumption about the structure of the output label space (in the form of hand engineered CRF pairwise potentials as in [6]). Or they rely on semantic maps from other sources to refine the initial predicted labels [5, 4]. However, these refinements through residual correction can only correct small errors (such as on the boundaries) while leaving major segmentation errors intact.

Built on these observations, we argue that we need to learn a joint structure of input and output variables that can effectively predict the replacements for major segmentation errors and thus, rectify them. Inspired from [10], we henceforth, propose a joint input-output model which segments the building footprints in high resolution imagery. While [10] addresses a continuous variable assignment task (disparity estimation), we solve here a binary classification problem. Our proposed method is buildup of two steps. Firstly, our model predicts an error map $E$ based upon the input image $X$ and initial segmentation $Y$. We then update the labels in regions of high error probabilities with a new label prediction which in turn relies on $X$, $Y$ and $E$ for its decision. Through the error map $E$, we learn a joint distribution between input and output variables which further helps in enforcing structure in the final label prediction. As shown in Figure 2, our model refines the blobby effect by learning the underlying structure of the building footprints which is enforced using the error maps. Finally, this leads our method to perform on par with existing state-of-the-arts on Inria Image Labeling dataset [1].

2. PROPOSED FRAMEWORK AND METHODOLOGY

Let’s assume our initial input image to be $X = x^{C \times H \times W}_{i=1}$, where $C$, $H$ and $W$ are the channel, height and width of $X$ respectively. Similarly, $Y = y^{1 \times H \times W}_{i=1}$, be the initial segmentation map. Our technique aims to model a joint relationship between input $X$ and output variables ($Y$) to rectify and produce a much more refined version of $Y$. This can be formulated as $Y' = F(X,Y)$ where $Y'$ denote the updated segmentation mask after replacing erroneous labels with new labels.

As shown in Figure 1, our proposed methodology comprises of two major steps. First, we predict errors ($E$) occurring in the initial building segmentation mask $Y$ with the help of input image $X$. Next, we utilize these error maps to correct those erroneous labels in $Y$, whose error probability is large enough (in $E$) by replacing them with predicted labels. In the following subsections, we explain these two steps in detail.

2.1. Error Detection

The error detection component ($F_e$) computes the probability map ($E$) to detect the erroneous labels in the initial segmentation mask $Y$, stated as:

$$E = F_e(X, Y).$$  (1)

In other words, $F_e$ learns from the joint input output space of $X$ & $Y$ to predict an error probability score map where each pixel has a value between 0 and 1. Specifically, it predicts whether or not a particular label $y_i$ is erroneous, if so, $y_i$ gets replaced with a correct label in the next step. $F_e$ can be easily formulated as a deep neural network which requires as such no explicit auxiliary loss and can be learnt under a single umbrella of one loss between corrected segmentation and ground truth segmentation.

2.2. Erroneous Label Replacement

The updated label $Y'$ is a convex combination of the initial segmentation mask $Y$ and updates from the replacement component denoted by $F_r$. It is given as:

$$Y' = E \odot F_r(X,Y,E) + (1 - E) \odot Y,$$  (2)

where $\odot$ represents element-wise product. The error map $E$ generated from $F_e$ acts as a gateway to restrict $F_r$ so as to just focus on the erroneous labels of $Y$ and replace them with the predicted ones. Similar to $F_e$, $F_r$ can also be modeled using any deep learning architecture. If we restrict the $E$ probability
maps to 0 and 1, the forward pass of Replacement happens as:

\[
Y' = \begin{cases} 
Y, & \text{if } F_e(X, Y) = 0 \\
F_r(X, Y, E), & \text{if } F_e(X, Y) = 1
\end{cases}
\]  

This shows that only erroneous labels are being replaced while non-erroneous labels remain intact. For any end loss function \( L \) between corrected segmentation mask \( Y' \) and ground truth labels, its gradient becomes:

\[
\frac{dL}{dF_r(\cdot)} = \begin{cases} 
0, & \text{if } F_e(X, Y) = 0 \\
\frac{dL}{dY'}, & \text{if } F_e(X, Y) = 1
\end{cases}
\]

In a way, the gradients update \( F_r \) only for those regions where erroneous labels are found in \( Y' \), thus restricting \( F_r \) to pay attention and predict replacements only for these particular regions. Additionally, passing Error maps to the \( F_r \) component, helps it to rely on the correct labels to predict replacements for the new erroneous labels. Altogether, \( F_r \) improves and makes correction of these erroneous labels, by jointly reflecting upon the Error maps, the input \( X \) and the initial segmentation \( Y \).

3. Training and Implementation Details

3.1. Dataset

We evaluated our method on the Inria Aerial Image Labeling Dataset [1] which consists of Airborne imagery of urban settlement over the United States and Austria. The entire dataset consists of two classes namely, building and not building. All the images are of size 5000 × 5000 and have a resolution of 30 cm with RGB bands. First 5 images from each class were chosen for validation, while the rest were used for training. For testing, we upload our test results to the project webpage.

3.2. Network Architecture

We initially train a Fully Convolutional Network [11] (FCN) adapted to Resnet-50 [12] architecture to generate our initial segmentation mask \( Y \). This model, which we name as Resnet50FCN, is trained to reconstruct ground truth segmentation masks with a given input image \( X \) and ground truth labels. We henceforth treat Resnet50FCN as our baseline.

For detection, \( F_e \) is implemented by using 5 convolutional layers (except last one, each is followed by batch-norm and Relu). The last conv. layer is followed by a soft-max, thus yielding us \( E \) between 0 and 1. To follow the input image size, we add an up-sampling layer on top of the Error map \( E \).

For replacement, \( F_r \) is implemented through compression block (compresses to 1/64 of the input resolution) and de-compression block (decompresses to 1/4 of input resolution). These are essentially residual blocks with parameterized skip connection between symmetric layers in decompression and compression blocks respectively.

For additional implementation details of detection and replacement modules, we refer the reader to Section 3.2 of [10].

3.3. Training

Using an L1 loss between ground truth and predicted output \( Y' \), we optimize using an adam solver [13], with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.99 \). The learning rate starting from \( 10^{-3} \) is decreases to \( 3 \times 10^{-4} \) at 12, \( 10^{-4} \) at 18, \( 3 \times 10^{-5} \) at 24 and finally \( 10^{-5} \) at 28 epochs. We continue training until 32 epochs. Each epoch consists of 500 batch iterations and each batch consists of 16 training samples where each sample is of size 1024 × 1024.

4. Experimental Results and Discussions

4.1. Quantitative Results

We show the quantitative results in Table 1 and Table 2 on both the validation and test dataset respectively, where we
Table 1: Evaluation results on validation set

<table>
<thead>
<tr>
<th>Method</th>
<th>Austin</th>
<th>Blooming</th>
<th>Inns.</th>
<th>S.Ps.</th>
<th>Tyrol-E</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-MLP [1]</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95.37</td>
<td>95.27</td>
<td>93.77</td>
<td>60.50</td>
<td>96.61</td>
<td>94.42</td>
</tr>
<tr>
<td>Multi-Task [9]</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>96.61</td>
<td>96.60</td>
<td>97.84</td>
<td>91.71</td>
<td>96.61</td>
<td>95.73</td>
</tr>
<tr>
<td>Resnet50FCN</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>88.12</td>
<td>97.02</td>
<td>99.3</td>
<td>89.07</td>
<td>95.12</td>
<td>96.87</td>
</tr>
<tr>
<td>ResnetDR</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>89.42</td>
<td>97.37</td>
<td>99.34</td>
<td>98.42</td>
<td>95.12</td>
<td>95.99</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results on test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Austin</th>
<th>Blooming</th>
<th>Inns.</th>
<th>S.Ps.</th>
<th>Tyrol-E</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-MLP [1]</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95.61</td>
<td>95.37</td>
<td>95.37</td>
<td>61.03</td>
<td>61.38</td>
<td>93.93</td>
</tr>
<tr>
<td>Resnet50FCN</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>63.34</td>
<td>95.00</td>
<td>96.61</td>
<td>78.07</td>
<td>91.67</td>
<td>72.07</td>
</tr>
<tr>
<td>ResnetDR</td>
<td>IoU</td>
<td>Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>64.27</td>
<td>95.99</td>
<td>96.52</td>
<td>71.70</td>
<td>78.68</td>
<td>72.07</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We present a novel technique for structured semantic segmentation of high resolution satellite imagery. To this end, we propose to learn the joint space of input-output variables. Subsequently, our method enforces structure in the form of predicting a much more regularized building footprint and hence, resolves to a large extent the problem of blobby effects as reported in the past methods. We compare our technique with the state-of-the-art methods on the Inria aerial image labeling dataset, where we perform on par with others.

6. REFERENCES