Automated Area Assessment of Objects Using Deep Learning Approach and Satellite Imagery Data

Kirill Tsyganov, Alexey Kozionov, Jaroslav Bologov, Alexandr Andreev, Oleg Mangutov and Ivan Gorokhov

Deloitte Analytics Institute, ZAO Deloitte & Touche CIS, Moscow, Russia,
{ktsyganov, akozionov, jbologov, aandreev, omangutov, igorokhov}@deloitte.ru

Abstract. We describe an actual case of applying deep neural networks for area assessment of different types of objects in selected geographical region through analysis of satellite images. The case was to detect, segment and asses area of buildings and agricultural lands on satellite images. We illustrate our framework of solving the problem and results validation methods. We compare performance of different convolutional neural networks in applying to our case and discuss the best quality segmentation model that was found – the U-net convolutional network. There was no training dataset of images and their corresponding masks available for our geographical region, but we constructed our own training set. Paper reports in detail on the processes of satellite imagery data preparation, images pre-processing, construction of training dataset and learning neural networks.

Keywords: Deep learning · Image segmentation · Object detection · Convolutional Neural Networks · U-net · Satellite imagery

1 Introduction

The paper presents main technical details of real-life client's case in experience of Deloitte Analytics Institute (Moscow). The paper does not pretend on scientific novelty of applied methods in the solution but rather describes our approach of using recent developments in machine learning in the actual industrial case.

Due to the existing country legislation, the client faced a lack of systematic recordings on agricultural and residential areas assessments and other national statistics. The client wanted to perform a structured audit of agricultural lands and residential areas paired with further monitoring of their development in time. The client requested us to provide a solution for an automated area assessment, based on an analysis of satellites imagery.

Since the problem required an accurate solution, we decided to use deep learning supervised approach. Basically we needed training dataset, neural network architecture for image segmentation and computational hardware resources to learn network on training data. We were going to experiment with publicly

available dataset [3] and in case of bad performance on test images of our region create own dataset for our region of interest. For the neural network architectures we took straightforward CNN [2] and more complex architecture with layers passing through each other [1]. For the networks' performance evaluation we took Jaccard index.

2 Satellite imagery data used in the solution

2.1 Data specific restrictions

In order to apply deep learning approach for image segmentation we needed training set of images, i.e. pairs of satellite images and their corresponding masks where only objects for detection were marked.

There was no training dataset with agricultural lands of out interest, so we had to construct our own dataset.

The geographical region of our research had specific desert environment and there was no training dataset of images for buildings segmentation of this region. To overcome this issue with labeled data we tried to use publicly available aerial imagery training set[1] of another geographical region(fig. 1). But test of models, trained on this open dataset, on images of our region of interest demonstrated insufficient quality of recognition. Possible causes of poor quality might be the following:

- due to the distinct geographical regions on the train and test images, buildings in the training dataset and buildings on the test images were very different: colors of roofs were different, shapes of buildings were different;
- projection angles on train and test images were different, it caused the size of shadows of objects;
- image color schemes on train and test sets were significantly distinct.

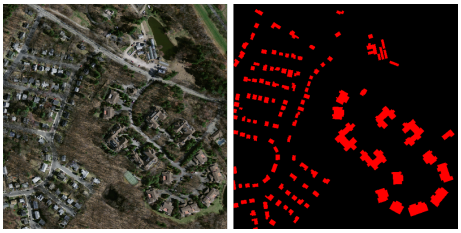

Fig. 1. Publicly available *Massachusetts Buildings Dataset* shared by V. Mnih [3]: consists of training dataset, i.e. satellite images and their corresponding masks with buildings

[1] *Massachusetts Buildings Dataset* publicly available at link `http://www.cs.toronto.edu/~vmnih/data/`.

2.2 Training dataset construction

After several unsuccessful attempts to use open training datasets for our problem, we came to the conclusion to use as training dataset satellite imagery data of our region of interest. Since there was no available labeled dataset we constructed such dataset by ourselves.

We used satellite images with resolution of 1 meter per pixel for training and test sets. Such resolution was able neural network to detect border structure of small buildings with area approximately 30 square meters.

To construct training dataset we took several small subregions and manually draw a mask with buildings and agricultural lands for it (fig. 2 and fig. 3). In order to improve generality power of out models we put in the training dataset buildings and agricultural lands of all types from different geographical subregions. Forming the training dataset was an iterative cycled process:

1. We trained model on the training dataset.
2. Then we tested model on test dataset.
3. Next we visually examined model's quality of recognition on test images and sought subregions where model performed low accuracy.
4. Finally we manually created masks for unsatisfactory recognized subregions and added such pairs of images-masks for the subregions into the training dataset.
5. Back to step 1.

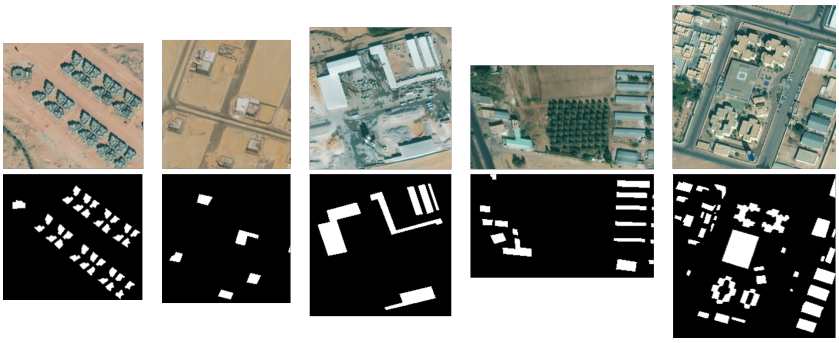

Fig. 2. Training dataset for buildings of our region marked by us: satellite images and their corresponding masks

2.3 Patches preparation

Due to purpose of fast training dataset formation, the images in the initial training dataset had the shapes of rectangles of different sizes. But the input for the neural network should have one predefined size. Therefore, in order to generalize our approach, for every image in the training dataset and its corresponding mask we took patches by sliding window of size 64×64 with step 16 (fig. 4).

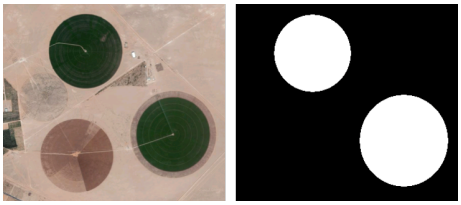

Fig. 3. Training dataset for agricultural lands of our region marked by us: satellite images and their corresponding masks

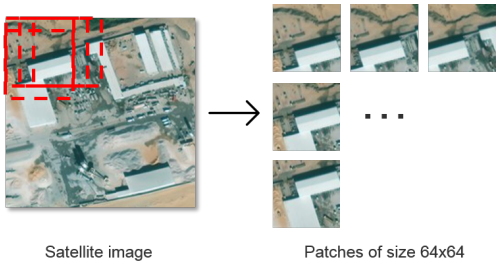

Fig. 4. Collecting patches from an image by sliding window

2.4 Image data augmentation

In order to enlarge training dataset without additional manual labelling of images we used standard techniques of image data augmentation, i.e. rotations and symmetries of original images (fig. 5). The data augmentation is applied to the patches of square shape, so that for every patch symmetry group of square is applied.

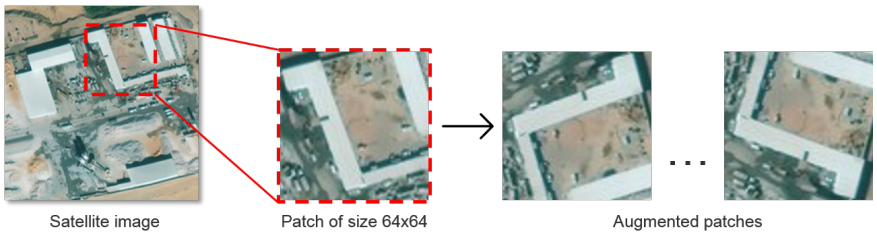

Fig. 5. Training dataset augmentation: first image is the original satellite image, next images are generated by augmentation

3 Evaluation metrics

Objects segmentation problems commonly estimated by the Jaccard index and visual analysis. We used the following metrics to assess the performance of models: Jaccard index, area error, precision, recall.

3.1 Jaccard index

The Jaccard index, also known as Intersection over Union (IU) is a measure of similarity and diversity of two sets. In order to compute Jaccard index between two finite sets A and B you need to divide the cardinality of intersection of A and B by the cardinality of union A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \quad 0 \leq J(A,B) \leq 1. \tag{1}$$

Jaccard index gives more penalty for error (both types of error) that precision and recall since it uses both false positives and false negatives statistics (fig. 6).

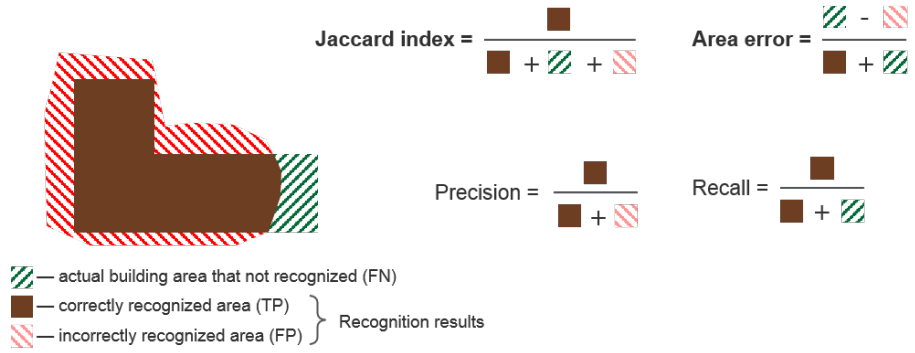

Fig. 6. Metrics used in the research

4 Solution

We had two classes of objects to detect and segment: buildings and agricultural lands with growing plants. Based on the conclusion that independent segmentation for multiple classes performs better than multinomial segmentation for multiple classes simultaneously [2], we decided to solve segmentation problems for each class separately. There was also an additional argument for such separation of problems – since second class of objects was agricultural lands with only growing plants we were going use additional features, like vegetation indexes [7] in order to increase accuracy of distinction of growing and not growing plants.

4.1 Buildings segmentation network architecture

We examined several architectures of convolutional neural networks: U-net [1] with different settings of hyperparameters and neural network with mixed convolutional and fully connected layers [2].

The architecture of the best performed CNN is based on U-net. Among other differences our network has less number of merging layers – 2 merges instead of 3 – we found that learning process CNN with 3 merges is very time consuming but does not give significant benefit in performance.

Our network (fig. 7) starts with contracting procedure with the repeated convolution, maxpooling and dropout layers and proceeds with expansion procedure in which maxpooling is substituted with upsampling. The most important and benefit feature in the network is the append of the output from the contracting layers to the input in the expansive layers. This approach significantly improves network performance on buildings' borders structure extraction. All convolutional layers except the last one use ReLU activation function and the output layer uses SoftMax.

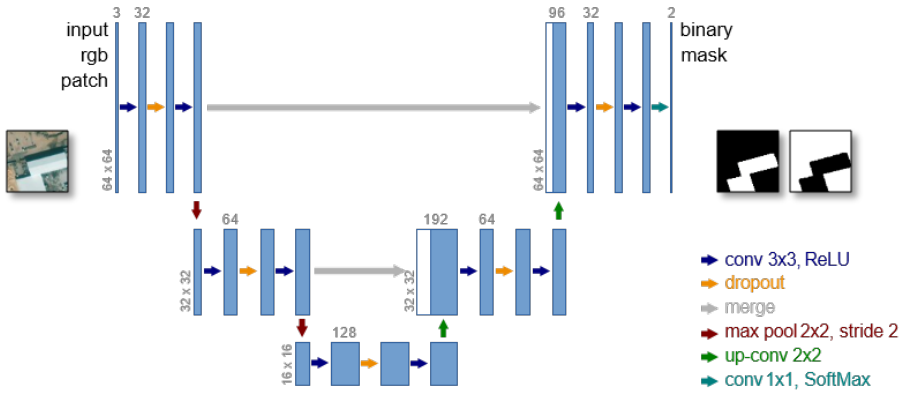

Fig. 7. CNN architecture for buildings segmentation based on U-net

4.2 Agricultural lands segmentation

In general the problem of lands segmentation is analogous to buildings segmentation. However, the average farm size is much bigger than average building size, so one need to cut initial image into considerably larger patches to preserve the information about farm structure and it's surroundings. The segmentation problem becomes computationally expensive when the neural network is used for processing heavy image patches.

A new approach was applied for circle farms recognition in order to overcome computational difficulties. The main feature of the approach is to use the combination of two heatmaps produced by different processing techniques to make

the final segmentation map. The first heatmap is produced by applying ellipsoid filters of various sizes to initial image. Exact sizes of the filters depend on image resolution. In this paper 5 x 5 and 50 x 50 filters were applied to 1 meter per pixel maps. Ellipsoid filter may be described as binary image of a circle inscribed in a square of a certain size or as matrix of zeros and ones with the ones filling the center circle-shaped region of the matrix. During applying of this filter erosion operation is performed. The filter slides through the image (like kernel in CNN convolution layer) and element-wise product of filter matrix and image segment is calculated. Minimum of these products is assigned to an anchor point that is set to be in the center of the filter. Thus, applying of a filter transforms initial image similarly to using convolution layer of CNN followed by minpooling layer. As a result, filtering, like CNN, is also produces a heatmap that is shown on Figure 9.

The second heatmap is produced by running random forest classifier which was trained to predict pixel class (farm / non-farm) based on it's color.

The idea behind proposed approach is to use the advantages of two techniques, which compensate each other flaws. Color segmentation method produce a relatively noisy heatmap, as the color of hills and roads is somewhat similar to farms color (especially when crop is not yet grown). Shape detection method—filtering—produces much less noise, but detected farms areas are significantly smaller than actual ones due to information loss during erosion process. In the joint heatmap calculated as average previous two the intensity of noise is lower than in color segmentation map and boundaries of farms are closer to actual than in shape detection map. Remaining noise can be removed by applying thresholding technique and median filter [5].

4.3 Polygons extraction

Neural network output due to the final softmax activation function provided us with two probabalistic heatmaps – one with probabilities of buildings and inverted one. But for the presentation results of recognition in the geospatial system it is necessary to convert heatmaps into polygons form. For this task we used thresholding of heatmaps and Douglas-Peucker algorithm [6].

5 Experiment results

We obtained Jaccard index of approximately 0.61 for buildings recognition, and 0.65 – for circle agricultural farms. The recognition results for buildings and circle agricultural farms can be seen at figures (8 and 9) correspondingly. As well as Jaccard Index, we computed the total area accuracy and it had value of 94% for buildings segmentation problem on the validation dataset.

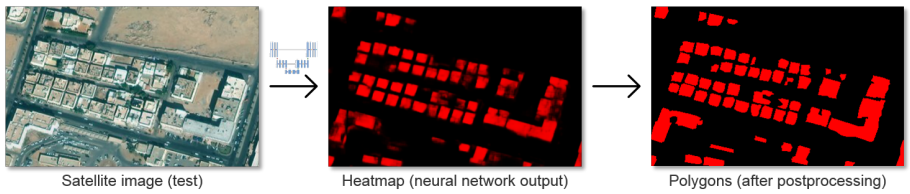

Fig. 8. Process of buildings' recognition on a test satellite image

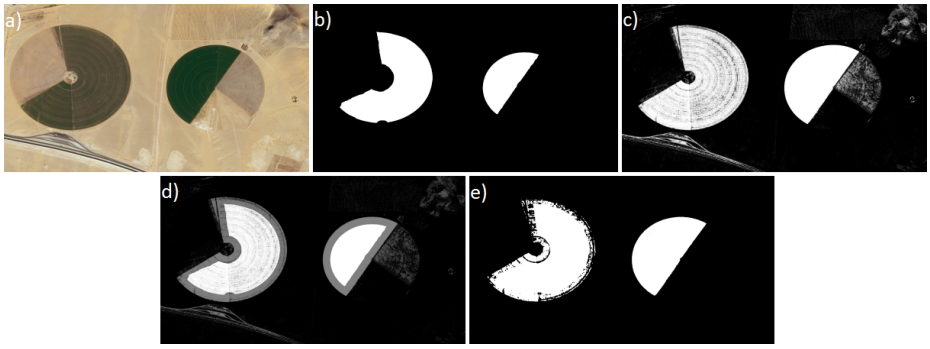

Fig. 9. a) initial satellite image, b) ellipsoid filters heatmap, c) color segmentation heatmap, d) joint heatmap, e) final heatmap after thresholding and applying median filter

5.1 Learning neural network for buildings segmentation

Since we have a binary classification problem (buildings, background) we used binary cross entropy as a loss function:

$$H_b(p) = H(p, 1-p) = -p\log(p) - (1-p)\log(1-p). \tag{2}$$

Learning process of unet with input and output patch of size 64×64 was not overfitting till approximately 85 epoch: starting from 85 epoch validation loss deviated significantly with training loss decreasing smoothly and it hurt the quality on test data (left picture on fig. 10).

Jaccard indices for different sizes of patches (as input and output shape for neural network) behaved the same starting from epoch 9 (right picture on fig. 10).

6 Discussion

We highlight the following branches of improvements that could be done for our solution:

– **Color histogram equalization of satellite images**
Since satellite images in the initial photo bank could be done by different satellites the color histograms of images can differentiate significantly.

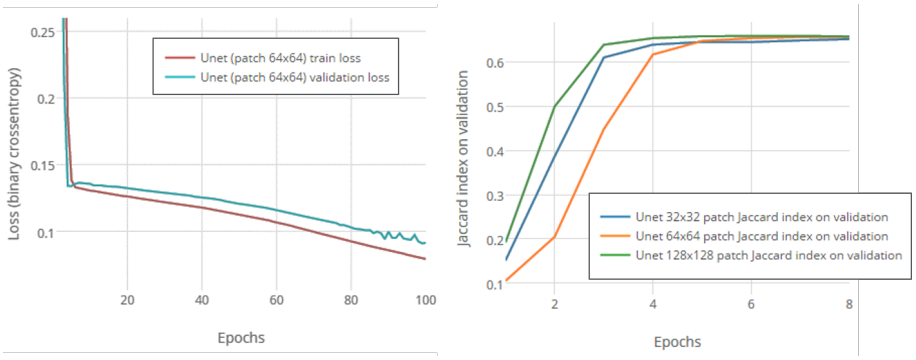

Fig. 10. Left: dynamics loss (binary cross entropy) during learning process of unet with input shape 64x64; right: Jaccard indices comparison for different patches during learning process

Such variety could harm the recognition quality. Therefore images' color histograms should be equalized before the further analysis. We suggest that contrast limited adaptive histogram equalization (CLAHE) [8] is the most appropriate method for images' color equalization.

– **Additional spectral bands**
Near-infrared range (NIR) and red edge channel could significantly enhance the quality of recognition algorithms, especially for agricultural lands. For example, combination of different bands with different resolution from different satellites in one regression model demonstrates high accuracy of agricultural land condition [9].
– **Training dataset formation**
Creating a mask for satellite image is a tough problem. In order to do a significant improvement of recognition's quality it is necessary to have masks for all types of objects a given class. We suggest to extend the training dataset not only by augmentation techniques of possessed images but by including bad-recognized regions.
– **Object detection phase**
Region proposal networks [10] [11] [12] [4] resolve the problem of object detection. The object detection phase can be used before image segmentation in order to reduce noise from other objects [4].
– **Object boundaries adjustment by probabilistic graphical models**
In order to improve localization accuracy of object boundaries it was proposed to use combination of methods from DCNNs and probabilistic graphical models [13]. Since CNNs can predict the rough position of the objects but it is difficult for them to highlight the boundaries, authors presented a new approach of refining objects' boundaries by applying fully-connected conditional random fields (CRF) for accurate boundary recovery after the final layer of the CNNs. They proved increased performance of this approach at PASCAL VOC-2012 image segmentation task so we think that the solution can be applied to our problem with benefit.

7 Conclusion

We present a report of applying deep learning approach for real life problem of objects' area assessment. We describe the whole solution process: collection of satellite imagery with appropriate resolution, creation of training dataset by manual labelling and data augmentations techniques, training and testing CNNs and extraction buildings' polygons from CNN's output heatmaps. We obtained the sufficient recognition quality (Jaccard index is 0.61 for buildings) with CNN based on U-net architecture [1]. Finally we propose the next steps of the recognition model design and feature engineering.

References

1. *Olaf Ronneberger, Philipp Fischer, and Thomas Brox*: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234–241, 2015.
2. *Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki*: Multiple Object Extraction from Aerial Imagery with Convolutional Neural Networks In: Journal of Imaging Science and Technology, Volume 60, Number 1, January 2016, pp. 10402-1-10402-9(9).
3. *V. Mnih*: Machine Learning for Aerial Image Labeling, Ph.D. thesis, University of Toronto, 2013.
4. *Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick*: Mask R-CNN. PAMI, 2017.
5. *J. W. Tukey*: Non-linear (non-superposable) methods for smoothing data, Int. Conf. Rec. 1974 EASCON, pp. 673.
6. *David Douglas, Thomas Peucker*: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, The Canadian Cartographer 10(2), 112–122, 1973.
7. *Rouse, J.W, Haas, R.H., Scheel, J.A., and Deering, D.W.*: Monitoring Vegetation Systems in the Great Plains with ERTS. Proceedings, 3rd Earth Resource Technology Satellite (ERTS) Symposium 1974, vol. 1, p. 309-313.
8. *K. Zuiderveld*: Contrast limited adaptive histogram equalization, Graphics gems IV, San Diego, CA:Academic Press Professional, Inc, 1994.
9. *Rasmus Houborg, Matthew F. McCabe*: High-Resolution NDVI from Planet's Constellation of Earth Observing Nano-Satellites: A New Data Source for Precision Agriculture, Remote Sens. 2016, 8, 768.
10. *Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik*: Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
11. *Ross Girshick*: Fast R-CNN, IEEE International Conference on Computer Vision (ICCV), 2015.
12. *Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun*: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Neural Information Processing Systems (NIPS) 2015.
13. *L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille* Semantic image segmentation with deep convolutional nets and fully connected CRFs, ICLR, 2015.