Forest Areas Segmentation on   
Aerial Images by Deep Learning

Vladimir Khryashchev  
P.G. Demidov Yaroslavl State UniversityYaroslavl, Russia  
v.khryashchev@uniyar.ac.ru

Anna Ostrovskaya  
People’s Friendship University of RussiaMoscow, Russia  
ostrovskaya\_aa@rudn.university

Vladimir Pavlov  
P.G. Demidov Yaroslavl State UniversityYaroslavl, Russia  
i@yajon.ru

Roman Larionov  
P.G. Demidov Yaroslavl State UniversityYaroslavl, Russia  
r.larionov@uniyar.ac.ru

Abstract — The aim of this research is to create a deep learning algorithm for automated forest areas segmentation on high-resolution aerial images. Loss values and Dice coefficient, which compares results of algorithms with real masks, was used to measure the quality of developed model. On the whole this paper demonstrates how convolutional neural network implemented on modern GPUs can be applied for the detection of forests on satellite images.

Keywords — computer vision, image segmentation, aerial images, forest area detection.

# Introduction

Remote sensing cannot completely replace ground-based data collection, but can help in monitoring large areas and hard-to-reach regions. Although computer vision algorithms for detecting objects in image develop quickly methods of automatic segmentation of satellite images is inferior to the man by the quality of work. Despite the rapid development of computer vision algorithms for detecting objects in an image, the task of segmentation of images of remote sensing of the Earth’s surface has not been brought to automatism with similar accuracy as with manual marking [1]. Although a human is able to solve a segmentation problem better than a computer, it takes too much time. In addition, in this case it is impossible to obtain results in real time, so the task of satellite image segmentation using computer vision algorithms is particularly relevant.

Today, large number of algorithms for detecting objects in an image exist. This problem is solved in the biometrics, medicine and robotics [2]. Most of these algorithms can be applied to remote sensing tasks. Segmentation of satellite images is a difficult task. The main approach of its solution based on machine learning methods is marking of the image pixels to corresponding classes of objects. Nowadays, the greatest effectiveness of solving this problem is achieved by using convolutional neural networks. The uniqueness of this method is based on the automatic determination of descriptors in the training process. So improvement of the segmentation accuracy and unique features that distinguish one class of objects from others is achieved [3].

Moreover, there are some reasons that CNN isn’t trivial solution of image segmentation. A unique approach is needed to solve the problem of the spatial extent of detected objects, taking into account the invariance to rotation of image or rescaling [4]. Such algorithms should [5, 6]:

* Have a sufficient number of sample images of each class in the training set. As a rule, open satellite image datasets don’t contain enough images. It’s necessary to expand the training sets of images by self-marking and mixing images from several datasets.
* Be invariant to rotation. Objects in the image can be rotated absolutely at any angle. The segmentation algorithm should be able to select the borders, regardless of the positioning of object in the image.
* Capture small spatial extent of objects. Most neural network algorithms solve the problem of selecting a large object in the image. These objects can be found in the ImageNet database [7]. Satellite images have a high resolution and cover large area where you need to find small, compared to the whole scene, objects. At the same time, if the work is not done with images of centimeters/pixel, most classes are deprived of unique small details that could become good distinguishing features of the class.

This paper discusses satellite imagery of the forest surface, which has the following features:

* high repeatability of shooting, thanks to which it is possible to repeatedly obtain data on the territory of interest, which increases the probability of obtaining cloudless or low-cloud images;
* large surface area with high spatial resolution;
* multichannel multispectral photos in the visible and invisible ranges, including ultraviolet and infrared channels

Information obtained in the process of shooting with RGB-channels, as well as with the near infrared channel (NIR), has a number of features in terms of their use in the analysis of forest areas.

The blue zone of spectral radiation is actively absorbed by chlorophyll (mainly chlorophyll B). This area is very sensitive to atmospheric conditions such as fog or haze. Compared with red or NIR channels, blue is less sensitive to changes in chlorophyll content. As a result, it is used only for special purposes, for example, water monitoring. To solve the problems of forestry, it’s best to use green and red channel composites to obtain high-quality color images that serve as the basis in geographic information systems. Blue channel facilitates recognition of forest fires in cloudless images [8].

Healthy vegetation mostly absorbs in red and blue spectrum, reflecting much of the green. The green channel serves not only to form a composite RGB image, but also allows humans to classify vegetation when it’s used in combination with other spectral channels. It’s also indispensable in assessing the overall condition of the forest

The red channel is very important for the analysis of vegetation (mainly forests) and is actively used. The wavelength of the red channel is greater than the blue one. For this reason, the state of the atmosphere affects it much less. The red channel plays a crucial role in the analysis of changes in forest cover, for example, in the mapping of damage from natural disasters, classification of vegetation species, monitoring of forest cover, etc. [9].

The reflectivity of tree foliage varies greatly in different species. The reflecting capacity of coniferous leaves is much lower than that of deciduous ones. Values (NDVIRE) NIR of young coniferous forest is higher than old one. Therefore, the NIR channel is very important for forest classification, determination of species composition, as well as for monitoring forest infestation by pests. The NIR channel also plays a key role in mapping the effects of hurricane winds, and is now becoming an important component in the calculation of some indicators that determine the biophysical parameters of vegetation.

This paper presents convolutional neural networks that can be used for forest segmentation. The training process, testing and special metrics for assessing the quality of neural network work are described.

# Neural Network Architecture

In this section we describe architecture of the neural network which was used for segmentation of forests in images taken by two different satellite sources.

The network is based on very popular and widespread architecture called U-Net, which is a convolutional neural network used for the task of semantic segmentation. Originally it was developed for segmentation of medical images of neuronal and other structures and outperformed many other competitors on the ISBI challenge [10].

U-Net is a u-shaped convolutional network, that is it consists of two parts: encoder and decoder. Both encoder and decoder are CNNs consisting of six blocks. Encoder’s each block includes two convolutions followed by rectified linear unit activation (ReLU) and max pooling operation. Encoder represents downsampling path. Decoder also consists of six blocks where each block includes up-convolution to upsample spatial size of each map, concatenation with corresponding feature map from downsampling path and two convolutions followed by ReLU. Decoder represents upsampling path which is used for restoration of segmentation mask. The last layer of the network is a convolutional K-channel layer, where K is the number of classes and its output is computed by applying pixel-wise softmax function. In our task, K is equal 2.

Since our images contain four channels: regular RGB and near-IR (NIR) channel, we modified U-Net by adding another encoder which separately accepts processes NIR channel. Results of both encoders are concatenated in the center of the network and also in each block of upsampling path like in original U-Net (see Fig. 1).

Modified version of the network was implemented in Python language using Tensorflow library. Tensorflow is a high performance graph-computation library leveraging GPUs for fast numerical computation and used for machine learning and deep learning tasks.



1. Architecture of U-Net network with 2 encoders



1. Examples of extracted RGB-patches and its corresponding masks.

# Numerical Results

It was necessary to perform some preprocessing steps before directly training the network. The whole dataset consists of 17 images in total from and each image channel contains 16-bit values unlike 8-bit RGB and also prone to the problem of outliers (or impulse noise). Presence of outliers negatively affects results and overall training time and convergence. So first, to tackle this problem we performed per-channel histogram equalization with min-max values chosen by thresholding cumulative distribution function of channel intensities (also known as so-called cumulative count cut). After this preprocessing step each image channel contains values in [0, 1] range with equalized histogram.

Since the dataset is very small, the second step was to split images into training and testing sets, which was performed by cutting each original image horizontally and taking part containing approximately 15% of objects’ pixels into test set, while taking the rest into training set.

The third step consisted in extracting patches from training and testing set as images sizes vary from 700 pixels to 12000 pixels per side and it would be impossible to train network using such data due to limited amount of GPU memory. Each patch represents image of size 512 by 512 pixels extracted from original image using sliding window with step 256 pixel per each axis. Examples of extracted RGB-patches and corresponding masks are shown at Fig. 2. Training set was also split into smaller training set and validation set to compute metrics to control training process per each epoch. Sizes of training and testing sets can be seen in Table I.

1. Sizes of training and testing sets

|  |  |
| --- | --- |
| **Training set** | |
| Total number of patches | 14139 |
| Number of patches with objects | 8655 |
| Number of patches without objects | 5484 |
| **Testing set** | |
| Total number of patches | 3282 |
| Number of patches with objects | 1785 |
| Number of patches without objects | 1497 |

It can also be seen that training dataset is imbalanced what negatively affects network’s learning process. To reduce this imbalance each batch is constructed by randomly selecting patches without objects and the same number of patches with objects. This way of constructing batch showed better segmentation and detection results in comparison to standard sequential feeding of images to network.

Moreover, to further increase size of training set we add different image augmentations: random flips (RF); rotations, spatial shifts, shifts in scale (SSR) and random noise in HSV color scheme (RN) which significantly increases quality of final segmentation.

Learning process is monitored after each epoch by evaluating loss and Dice coefficient computed on validation set. Dice coefficient:

where X and Y are grayscale or binary masks. We used two losses to train models:

1. Binary cross entropy loss + dice loss (BCE + DL):



where

is the label of sample , is the predicted probability of sample and is the Dice coefficient.

2. Focal loss + dice loss (FL + DL):

where

is the focal loss and .

Focal loss was specifically developed for tasks where the data is highly imbalanced and outperformed other competitors on COCO dataset.

We trained the following four models (BB is for balanced batch) during 100 epochs with batch size of 16 images and Adam optimizer:

1. BCE + DL / BB / RF + SSR (BCE #1)

2. BCE + DL / BB / RF + SSR + RN (BCE #2)

3. FL + DL / RF + SSR (FL #1)

4. FL + DL / RF + SSR + HSV (FL #2)

Models were trained on supercomputer NVIDIA DGX-1 in AI-center of P.G. Demidov Yaroslavl State University. Loss values and Dice coefficient values changes during validation process are shown in Fig. 3.

The final network output is the segmentation map with values ranging from 0 to 1 and it is necessary to select appropriate value for threshold to obtain binary mask. For this task we used simple grid search to maximize F1 on validation set with step of . Selected values of threshold for different models are shown in Table II.

1. Selected values for threshold on validation set

|  |  |  |
| --- | --- | --- |
| **Model** | **T** | **F1** |
| BCE #1 | 0.39 | 0.3089 |
| **BCE #2** | **0.62** | **0.4504** |
| FL #1 | 0.52 | 0.2175 |
| FL #2 | 0.56 | 0.361 |

Results obtained on test set are shown in Table III. F1 is the standard F-measure computed using precision P and recall R, both of which are calculated according to correct detection. Object is considered correctly detected if it has with ground truth object. It can be seen that best results showed model BCE #2 for . FL #2 with RN did not show any significant increase in quality of segmentation. Moreover, it can be supposed that it produces too many false positives as P value is quite low. The same can be said about BCE #1.

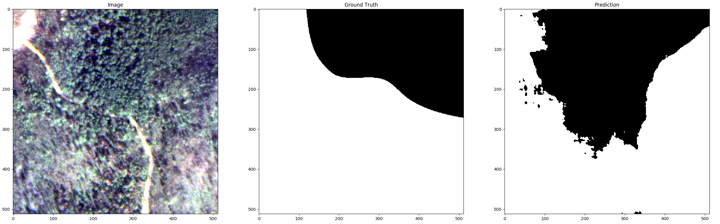
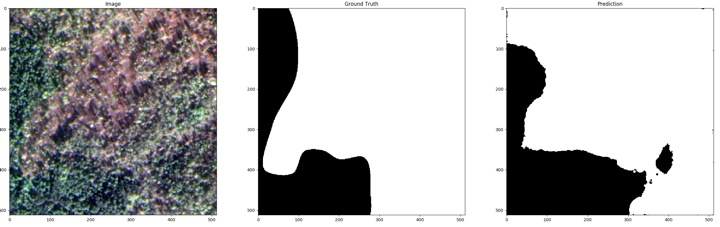
1. Segmentation results on test set for different values of threshold

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **T** | **D** | **F1** | **P** | **R** |
| BCE #1 | 0.39 | 0.8439 | 0.1641 | 0.09475 | 0.6136 |
| BCE #1 | 0.5 | 0.8446 | 0.17 | 0.09873 | 0.6105 |
| **BCE #2** | **0.62** | **0.7652** | **0.3488** | **0.248** | **0.5876** |
| BCE #2 | 0.5 | 0.7554 | 0.3319 | 0.2309 | 0.59 |
| FL #1 | 0.52 | 0.8222 | 0.1967 | 0.1168 | 0.6211 |
| FL #1 | 0.5 | 0.8181 | 0.1775 | 0.1038 | 0.6144 |
| FL #2 | 0.56 | 0.7874 | 0.1549 | 0.09027 | 0.5443 |
| FL #2 | 0.5 | 0.7632 | 0.1288 | 0.07269 | 0.5664 |

Examples of final segmentation are shown in Fig. 4. The first image is source patch, the second image is ground truth mask and the third one is algorithm prediction.



1. Loss values and Dice coefficient on computed validation set

1. Examples of aerial forest areas images segmentation

# Conclusion

In this article we showed that our proposed modification of U-Net with balanced batch is capable of segmenting forests in satellite images and generalization to test set. Although it does not show high results in detection (max value of F1 is 0.34), it can be connected with incorrect or incomplete segmentation of ground truth images by experts, since visually prediction on many test examples look correct. Further research is required to increase quality of segmentation and detection what can be done by increasing number of source images in dataset, adding more aggressive augmentations and usage of modified versions of U-Net with new losses that can correctly tackle problem of imbalanced data.

##### Acknowledgments

This article was prepared with the financial support of the Ministry of science and education of the Russian Federation under the agreement No. 075-15-2019-249 from 04.06.2019 (identifier works RFMEFI57517X0167).

##### References

1. Sayfeddin D., Bulgkov A., Kruglova T. Neirosetevaya sistema otslejivaniya mestopolozheniya dinamicheskogo agenta na baze kvadrokoptera // Inzhenernyi vestnik Dona, 2014, No. 1, URL: http://ivdon.ru/magazine/archive/n1y2014/2293/
2. Plutogarenko N., Varnavskiy A. Primenenie neironnykh setei dlya postroeniya modeli prognozirovaniya sostoyaniya gorodskoi vozdushnoi sredy // Inzhenernyi vestnik Dona, 2012, No. 4-2, URL: http://ivdon.ru/magazine/archive/n4p2y2012/1351/
3. L. Deng, D. Yu Deep Learning : Methods and Application Foundations and Trends in Signal Processing. 2014, vol. 7, no. 3-4, pp. 197-387.
4. Solov’ev R. A., Telpukhov D. V., Kustov A. G., Avtomaticheskaya segmentatsiya sputnikovykh snimkov na baze modifitsirovannoi svertochnoi neironnoi seti UNET (Automatic segmentation of satellite images based on the modified UNET convolutional neural network), Inzhenernyi vestnik Dona, 2017, Vol. 47, No. 4(47), URL: http://ivdon.ru/uploads/article/pdf/IVD\_56\_soloviev\_N.pdf\_116222c2f5.pdf.
5. V.Khryashchev, L.Ivanovsky, V.Pavlov, A.Ostrovskaya, A.Rubtsov Comparison of Different Convolutional Neural Network Architectures for Satellite Image Segmentation // Proceedings of the FRUCT’23, Bologna, Italy, 13–16 November 2018. pp.172–179.
6. Khryashchev V., Pavlov V., Priorov A., Ostrovskaya A. Deep learning for region detection in high-resolution aerial images // IEEE East-West Design & Test Symposium (EWDTS). 2018. pp. 1–5.
7. ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), Web: http://image-net.org/challenges/LSVRC/2012/
8. W G.. Rees Physical principles of Remote Sensing. Cambridge University Press, 2006, 336 p.
9. Campell J.B. Introduction to remote sensing / J.B. Campell.–N.Y.–London: The Guilford press, 1996–P. 120–549.
10. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, vol. 9351, 2015, pp. 234–241.