# Optimization Of Convolutional Neural Network For Object Recognition On Satellite Images

Vladimir V. Khryashchev, Vladimir A. Pavlov,
P.G. Demidov Yaroslavl State University,
Sovetskaya 14, Yaroslavl, Russia
v.khryashchev@uniyar.ac.ru, i@yajon.ru

Anna A. Ostrovskaya, Alexander S. Semenov
People's Friendship University of Russia, 6
Miklukho-Maklaya st., 117198, Moscow, Russia
ostrovskaya_aa@rudn.university,
semenov.venture@mail.ru

***Abstract*—Investigation about using convolutional neural networks for detection geo-objects on the satellite images from Landsat-8 was presented. U-NET convolutional neural network architecture for implementing the recognition algorithm was used. The neural network was trained by the marked image base "Urban Atlas". Urban Atlas contains images of 21 classes, but in current research was used 3 classes: "Forest", "Agriculture" and "Water". Images obtained from the Landsat-8 satellites are used for estimation of automatic object detection quality. To analyze the accuracy of the object detection algorithm, the selected regions were compared with the areas by previously marked by experts. Two modification of object detector were proposed and analyzing.**

***Keywords — image recognition, convolutional neural network, satellite images, object detection, deep learning.***

## 1. INTRODUCTION

In machine learning applications, aerial image interpretation is usually formulated as a pixel labeling task. Given an aerial image like the goal is to produce either a complete semantic segmentation of the image into classes such as building, road, tree, grass, and water [1, 2] or a binary classification of the image for a single object class [3-5].

Convolutional neural networks have become ubiquitous in computer vision ever since AlexNet [6] popularized deep convolutional neural networks by winning the ImageNet Challenge: ILSVRC 2012 [7]. An example of implementing a convolutional neural network for land-use classifying (detection of ravines, forest areas, etc.) can be found in [8]. This network was trained on a publicly available UC Merced Land Use Dataset [9] (2100 images of 21 classes). Classification of agricultural territories is presented in the article [10]. The authors of this study adapted the architecture of the VGG16 neural network [11], which is preliminarily trained to recognize images of the ImageNet dataset.

Object detection is a common task in computer vision, and refers to the determination of the presence or absence of specific features in image data. Once features are detected, an object can be further classified as belonging to one of a pre-defined set of classes. This latter operation is known as object classification. Object detection and classification are fundamental building blocks of artificial intelligence. A major challenge with the integration of artificial intelligence and machine learning in aerial image analysis is that these tasks are not executable in real-time or near-real-time due to the complexities of these tasks and their computational costs. One of the proposed solutions is the implementation of a deep learning-based software which uses a convolutional neural network algorithm to track, detect, and classify objects from raw data in real time. In the last few years, deep convolutional neural networks have shown to be a reliable approach for image object detection and classification due to their relatively high accuracy and speed.

Remote sensing data and deep learning methods are also used for solving the problems related to the object detection task. For example, in [12] the geo-positioning of the photograph is determined by the visual correlation of the image content and the satellite image using convolutional neural networks. Other applications in this area address poverty prediction in some African countries. For this purpose, night-time images are used for assessing the ratio of city illumination, as evidence of economic activity [13].

The European Urban Atlas [14] provides reliable, inter-comparable, high-resolution land use maps for 305 Large Urban Zones and their surroundings (more than 100.000 inhabitants as defined by the Urban Audit) for the reference year 2006 in EU member states and for 695 Functional Urban Area (FUA) and their surroundings (more than 50.000 inhabitants) for the reference year 2012 in EU and EFTA countries. Change layers were produced in 2012 and only for all FUAs covered both in 2006 and 2012 reference years.

Our work is devoted to the analysis of the use of convolutional neural networks for detecting earth surface types using remote sensing data. For deep learning of convolutional neural networks we used the marked image database UrbanAtlas. Urban Atlas contains images of 21 classes, but in current research we use 3 classes: "Forest", "Agriculture" and "Water". Images obtained from the Landsat-8 satellites [15] are used for estimation of automatic object detection quality.

## 2. THE NEURAL NETWORK LEARNING PROCESS

The training of the neural network for the detection of objects on satellite images we used 500 unique images. Each satellite image is downloaded from the Landsat 8 and includes an area of one square kilometer from summer 2017. Prepared images were divided into 2 datasets: training and testing with a proportion of 70% to 30%.

We need to prepare a dataset to train the U-NET neural network, where each pixel is assigned a specific class. We chose the training images were each class was balanced. The percentage of each of three class in the training step is presented in Fig. 1.

Jaccard index was chosen as the optimization parameter for learning the neural network, which will determine the degree of similarity of the compared objects. The index is determined by the formula (1):

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

range of values $0\le J(A,B)\le 1$

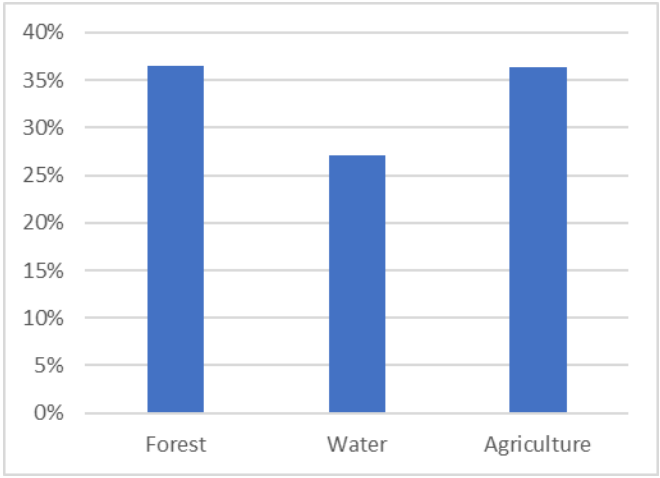
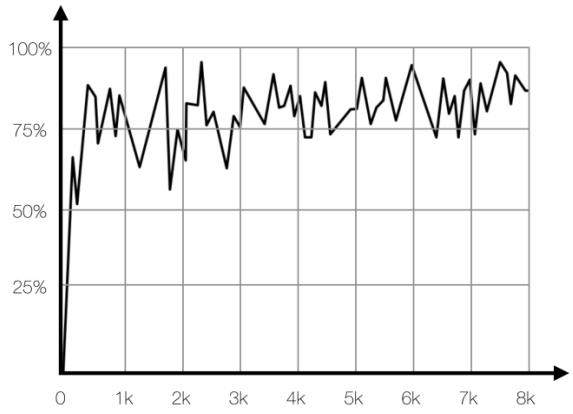
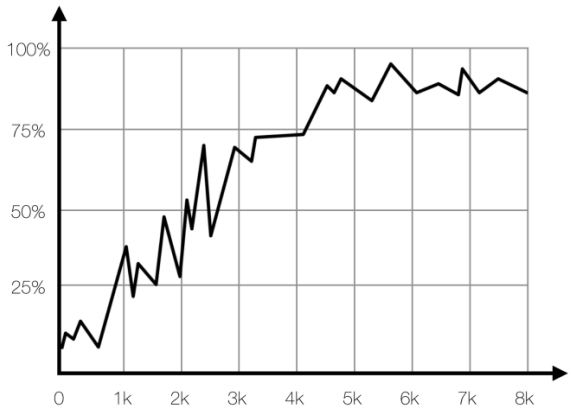



Fig. 1. The ratio of data examples to each class in the training set




a)




b)

Fig. 2. Accuracy of "water" class: detection during training (a) and during verification (b).

If the index of the coefficient tends to 1, then the geometric objects are identical, if 0 it means that objects not equal. Training of the neural network was carried out in 60 epochs. This approach has made it possible to avoid re-training and to obtain the maximum detection accuracy for each detector. Each epoch was trained on 450 batches. Each batch includes 200 parts of images that are randomly allocated from the training dataset. Each part is an image of 224x224 pixels. This image size proved to be the most optimal, since we achieved a high learning rate, and for these detection classes there is no need to distinguish very small features. Increasing the size of a part of the image will be relevant when adding the following detection classes, such as transport, small building objects, and airplanes.

Fig. 2 shows the accuracy of the neural network detection during training (a) and the test on the test dataset (b).

The loss function is shown in Fig. 3.

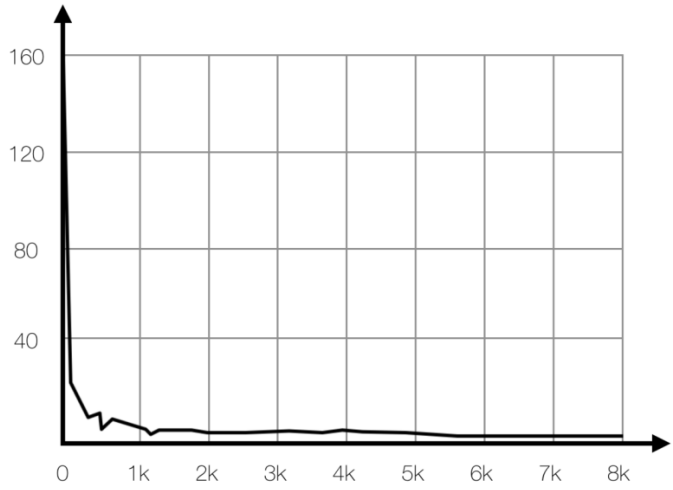



Fig. 3. The graph of the loss function for learning the neural network architecture of U-NET for class water resources.

## 3. OPTIMIZATION OF THE DETECTOR

The classical detection approach contains images of the same training parts are fed to the input of a trained neural network with the U-NET architecture. Within the proposed detector, these are satellite images of 224x224. Since the detection rate of this network is low, for the highest speed of the satellite image bypass, the detector is sequentially cutting the image, as shown in Fig. 4.

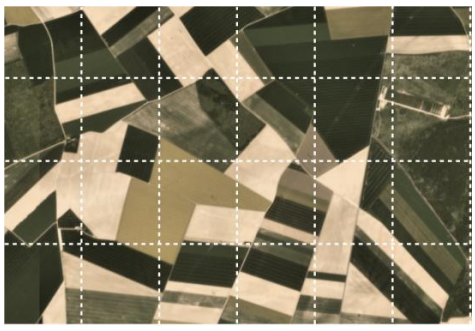


Fig. 4. Separation of the image into parts for feeding the detector

As shown by the results of experiments, the extreme pixels of the input image of U-NET detect with the least accuracy than the central ones. As a result, optimization of the image mask bypass was performed. Each fragment is traversed 2 times with an offset of ½ image (Fig. 5).

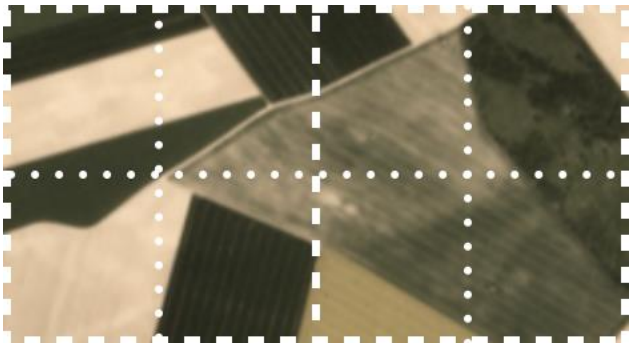


Fig. 5. Example of bypass of 2 cells with displacement

The probabilities of intersecting parts are defined as the arithmetic mean of the probabilities of each class. The resulting value of the resulting class is determined by the highest average percentage of the probability of detection.

$$R = \max \frac{\sum_{i=1}^{N} P_i}{N} \tag{2}$$

The result of the proposed detector is the assignment to each pixel of the image of the detection class (forest, water, agriculture). For productive work with this information it is necessary to obtain polygons of the detected data. We solved the task of clustering for this. We compared the work of two clustering algorithms: k-means and discretize, which are included in the sklearn python library. Fig. 6 shows the results of these clustering algorithms.

It can be seen that k-means most accurately determines the outer boundaries of the detection zone, but discretize greatly reduces the detection error itself, making objects more geometric. As a result, in the detection system, we applied the discretize algorithm with default parameters.

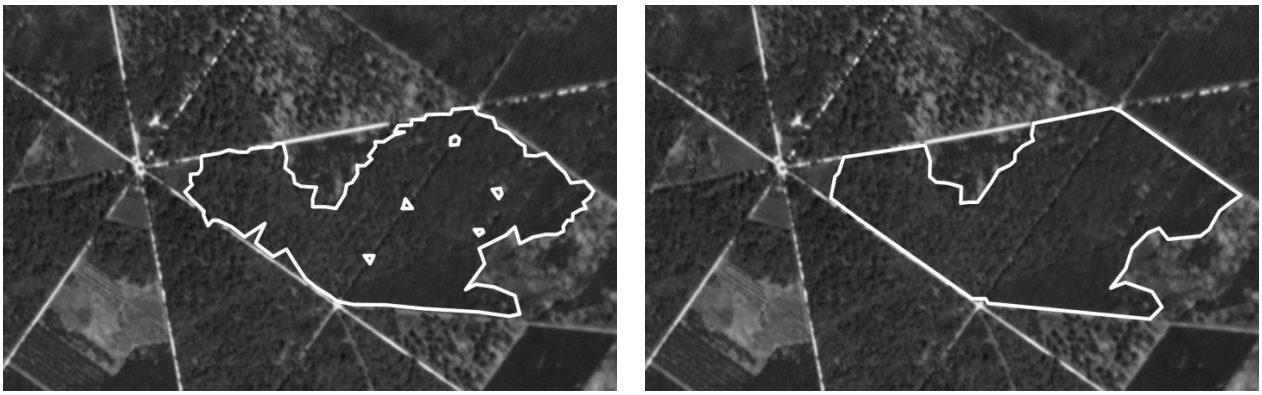
## 4. EXPERIMENTAL RESULTS

To analyze the accuracy of the object detection algorithm, we compared the contours of automatically detected areas with areas of expert markup. To analyze the accuracy of the object detection algorithm, the selected regions were compared with the areas by previously marked by experts. For the experiment we selected 100 space images where there is an interest class. Then the procedure of automatic detection of the object in the image was carried out. Next we made detection of objects by our detector on neural networks. The resulting multipoligons we compared to the percentage of intersection.

$$R = \frac{Spred \cap Sexp}{Sexp} \tag{3}$$

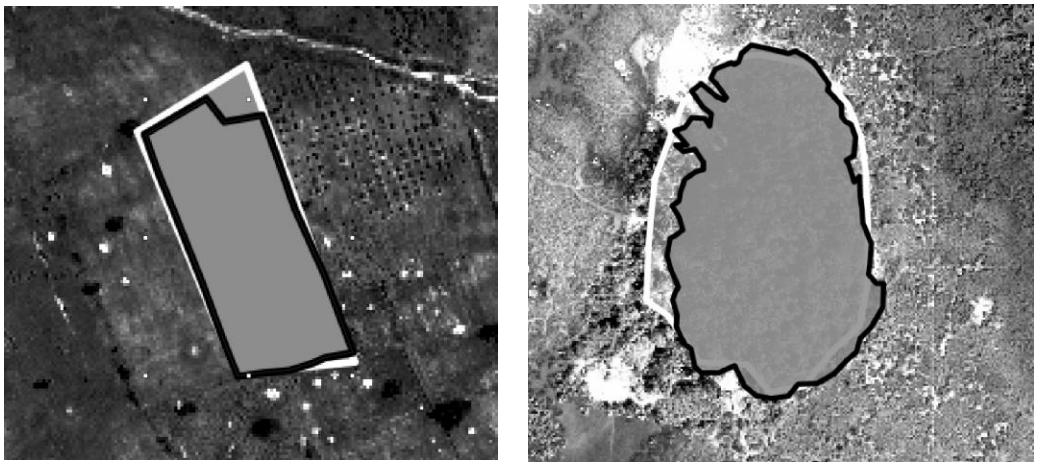
where S*pred* – detected multipolygon, S*exp* – expert multipoligon.

Examples of images with detected and expert multipolygon is shown on Fig. 7. A white solid line identifies the contour, drawn by the detector, a black solid line is a multipoligon, drawn by an expert.



a) b)

Fig. 6. Example of clustering of data detection for obtaining a multipolygon a) k-means, b) discretize



a) b)

Fig. 7. Images from Landsat-8 with automatically (white curve) and manual (black curve) marked region: a) "agriculture", b) "forest"

We created 2 detectors to confirm our hypothesis of the choice of loss function. The results of the experiments are given at Table 1 and Table 2. In this tables total objects – is the total number of expert markings of multipoligons of this class and detection accuracy calculated by the formula (4)

$$Accuracy = \frac{True\ detection}{True\ detection + False\ detection} \quad (4)$$

TABLE 1. OBJECT DETECTION RESULTS BY SIMPLE DETECTOR

| | Total objects | True detection | False detection | Detection accuracy | Average percentage of intersection |
|---|---|---|---|---|---|
| Forest | 200 | 160 | 24 | 86,96% | 89,62% |
| Water | 40 | 33 | 6 | 84,62% | 80,83% |
| Agriculture | 220 | 199 | 30 | 86,90% | 91,34% |
| | | | | ~ 86,16% | ~ 87,26% |

TABLE 2. OBJECT DETECTION RESULTS BY IMPROVED DETECTOR

| | Total objects | True detection | False detection | Detection accuracy | Average percentage of intersection |
|---|---|---|---|---|---|
| Forest | 200 | 168 | 21 | 88,89% | 92,32% |
| Water | 40 | 38 | 4 | 90,48% | 81,71% |
| Agriculture | 220 | 208 | 11 | 94,98% | 96,17% |
| | | | | ~ 92,73% | ~ 90,07% |

From Table 2, the highest percentage of intersections of areas of detected objects of the "Agriculture" class. This is due to the clarity of the boundaries and the apparent visual separation of the surrounding objects. When we detect “Water”, we have a lot separate parts of the water resource are allocated. This is due to the presence of ice and other objects over rivers and lakes. Forest territory has an average detection value of 92.3% of the intersection of areas due to inaccurate allocation of boundaries of forest territory.

## 5. CONCLUSIONS

The final classification accuracy is 81.7% for objects such as "water resource", 92.3% for objects of the "forest" class and 96.1% for objects of the "agriculture" class. The considered algorithm can be applied for the semantic analysis of images from the satellite: allocation of the territories of cities, control of construction and other.

## 6. ACKNOWLEDGMENT

The paper was prepared with the financial support of the Ministry of Education of the Russian Federation in the framework of the scientific project No. 14.575.21.0167 connected with the implementation of applied scientific research on the following topic: «Development of applied solutions for processing and integration of large volumes of diverse operational, retrospective and the thematic data of Earth's remote sensing in the unified geospace using smart digital technologies and artificial intelligence» (identifier RFMEFI57517X0167).

The authors are grateful to AI-center of P.G. Demidov Yaroslavl State University for providing access to the supercomputer NVIDIA DGX-1.

REFERENCES


[1] S. Kluckner and H. Bischof, “Semantic classication by covariance descriptors within a randomized forest”, In Computer Vision Workshops (ICCV), pp. 665-672, 2009.

[2] S. Kluckner, T. Mauthner, P. M. Roth, and Horst Bischof, “Semantic classication in aerial imagery by integrating appearance and height information”, In ACCV, volume 5995 of Lecture Notes in Computer Science, pp. 477-488. Springer, 2009.

[3] V. Mnih and G. Hinton, “Learning to detect roads in high-resolution aerial images”, In Proceedings of the 11th European Conference on Computer Vision (ECCV), September 2010.

[4] V. Mnih and G. Hinton, “Learning to label aerial images from noisy data”, In Proceedings of the 29th Annual International Conference on Machine Learning (ICML 2012), June 2012.

[5] P. Dollar, Z. Tu, and S. Belongie, “Supervised learning of edges and object boundaries”, In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1964-1971, 2006.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, In Advances in neural information processing systems, pp. 1097–1105, 2012.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge”, International Journal of Computer Vision, 115(3), pp. 211–252, 2015.

[8] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, “Land Use Classication in Remote Sensing Images by Convolutional Neural Networks”, Web: https://arxiv.org/abs/1508.00092, 2015.

[9] Y.Yang and S. Newsam, “Bag-of-visual-words and Spatial Extensions for Land-use Classication”, In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS ’10). ACM, NY, USA, pp. 270–279.

[10] O. A. B. Penai, K. Nogueira, and J. A. dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?”, In 2015 IEEE Conference on Computer Vision and Paern Recognition Workshops (CVPRW), pp. 44–51.

[11] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, Web: https://arxiv.org/abs/1409.1556 (2014).

[12] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, “Predicting Ground-Level Scene Layout from Aerial Imager”, Web: https://arxiv.org/abs/1612.02709 (2016).

[13] N. Jean, M. Burke, M. Xie, W.M. Davis, D.B. Lobell, and S.Ermon, “Combining satellite imagery and machine learning to predict poverty”, Science 353, 6301 (2016), pp. 790–794.

[14] “European Union. 2011. Urban Atlas”, Web: https://www.eea.europa.eu/data-and-maps/data/urban-atlas.

[15] “Landsat8”, Web: https://en.wikipedia.org/wiki/Landsat_8.