High-resolution Aerial Image Segmentation for Automated Building Detection

Leonid Ivanovsky [1)], Anna Ostrovskaya [1)], Vladimir Khryashchev [2)], Anatoly Sedov [2)]
1) People's Friendship University of Russia (RUDN University), 117198 Moscow, Russian Federation, Miklukho-Maklaya str.6, e-mail: leon.ivanovsky@yahoo.com ostrovskaya_aa@rudn.university

2) P.G. Demidov Yaroslavl State University, 150003 Yaroslavl, Russian Federation, Sovetskaya str. 14, email: v.khryashchev@uniyar.ac.ru, agsedov@gmail.com

Abstract: *The goal of our research was to develop convolutional neural network methods for automatically extracting the locations of buildings from aerial images. To analyze the quality of developed algorithms, there was used Sorensen-Dice coefficient of similarity which compares results of algorithms with real masks which were generated from json files. All in all we show how deep neural networks implemented on modern GPUs can be used to efficiently learn and detect needed objects. The problem of building detection on satellite images can be put into practice for urban planning, building control, etc.*

Keywords**:** computer vision, deep learning algorithms, satellite image segmentation, building detection.

1. INTRODUCTION

Nowadays, the problem of object detection on satellite images is in the focus of researchers. Automatic image segmentation allows to extract areas of interest such as vehicles or buildings. Most approaches of solving this problem suggest the development of deep learning algorithms.

In machine learning the problem of image segmentation is usually reformulated as a classification of each pixel. The simplest and the slowest way of solving this problem is a manual segmentation of images by experts. However, it is a time-consuming process, which is subjected to human errors. Automatic image segmentation. makes possible to process images immediately after receiving it. Satellite image segmentation finds its application in urban planning, building control forest management and meteorology.

This article presents developed convolutional neural networks (CNNs). The main advantage of CNNs is that they can detect and classify objects in real time while being computationally less expensive and superior in performance when compared with other machine learning methods. Essentially, the mathematical structure of CNNs is parallel and perfectly fits the architecture of graphics processing units (GPUs) which consists of thousands of cores to perform several tasks simultaneously [1]. CNNs have become ubiquitous in computer vision since AlexNet [2] won the ImageNet Challenge: ILSVRC 2012 [3].

The problem of satellite images segmentation is challenging. The most approaches of solving this problem suppose the usage of deep learning algorithms. The features in these networks are formed automatically in the process of training. In recent years there were developed some CNNs, which aim at satellite image segmentation.

One of the most successful algorithms is based on pyramid architecture of CNNs - Feature Pyramid neural networks (FPNs). FPNs showed acceptable results of object detection on satellite images [4]. The usage of FPNs allows to get the value of Jaccard index is approximately equal to 0.49 for satellite images from DeepGlobe [5].

In paper [6] there is presented U-Net architecture – a specific type of FPN, which had received a lot of interest for segmentation of biomedical images. Later this model was applied to pixel-wise classification of satellite images [7]. U-Net uses skip-connections to combine low-level and higher-level maps of features. Using U-Net architecture, the authors of paper get the value of Sorensen-Dice coefficient is equal to 0.75 for building detection on satellite images.

In paper [8] authors present LinkNet. This special architecture of CNNs, which consist of an encoder and a decoder as U-Net. It efficiently share the information learnt by the encoder with the decoder after each downsampling block. This technique of feature transfer allows to get high accuracy values for object detection on the CamVid dataset [9].

This article consists of six parts. The first part is devoted to CNNs as an approach in machine learning and peculiarities of satellite images segmentation. It also contains an overview of some papers for object detection on aerial photos. The second part is devoted to the available databases of satellite images. The third section describes the developed architectures of CNNs for building detection on aerial photos and some peculiarities of training of models. Futhermore, in this part there is mentioned about Keras framework and Tensorflow library. The forth part presents the results of numerical experiments for the developed models. In the conclusion there is summarized the research. And finally, the last section represents references.

2. DATABASES OF SATELLITE IMAGES

A standard database of images is the important part for learning, efficiency evaluation and comparative analysis of different machine learning algorithms. Nowadays, there are some available databasets of satellite images.

The IKONOS database [10] contains 11-bit and 8-bit color stereo images from IKONOS satellite in GEOTIFF format. The main attributes of these images are 360-degree pointing capability and a base-to-height ratio of 0.6. Every image of IKONOS database has a spatial resolution of 0.82 m / pixel and was shooted with an elevation angle from ground to sun more than 15° with less than 15%

cloud cover. Examples of images from the IKONOS database are shown in Fig. 1.

Fig. 1. - Examples of images from the IKONOS database

The Quickbird database [11] contains 11-bit color images from Quickbird satellite. On January 27, 2015, one of DigitalGlobe's [12] oldest and most historically significant satellites re-entered Earth's atmosphere after completing its 13-year mission in orbit. QuickBird made more than 70,000 trips around the planet, capturing some 636 million square kilometers of high-resolution. Each image of Quickbird database has a spatial resolution of 0.6 m / pixel. Examples of images from the Quickbird database are shown in Fig. 2.

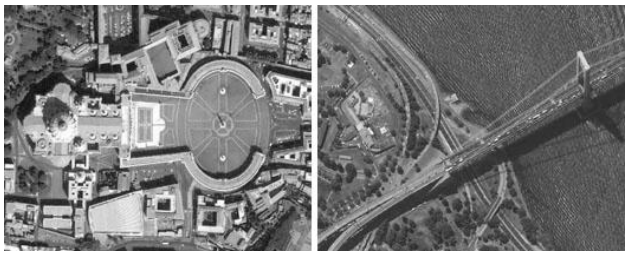

Fig. 2. - Examples of images from the Quickbird database

DSTL dataset contains 50 satellite images in GEOTIFF format. This database was provided in Kaggle competition "DSTL Satellite Imagery Feature Detection" [13]. Images of DSTL dataset are labeled on 10 different classes: "buildings", "manmade structures", "roads", "tracks", "trees", "crops", "waterway", "standing water", "large vehicles" (e.g. lorries, trucks or buses) and "small vehicles" (cars, vans or bikes). The examples of images from the DSTL database are shown on Fig. 3.

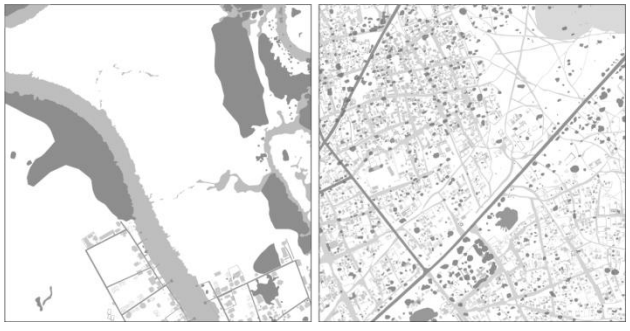

Fig. 3. - Examples of images from the DSTL database

In our research there were used 14 color satellite images from the private database of Russian cities. Every image of this dataset has a resolution of 16384 × 16384 pixels with a spatial resolution of 0.5 m / pixel. Images of private database cover the area of three Russian cities: Yaroslavl, Rybinsk and Moscow. Examples of images from the private database are shown in Fig. 4.

Fig. 4. - Examples of images from the private database

3. CONVOLUTIONAL NEURAL NETWORKS

This article presents developed CNNs. This type of neural networks has a special architecture, aimed at quick and high-quality segmentation [14]. In this research there is made a comparative analysis of U-Net [7] и LinkNet [8]. The research of work of developed models continues the research, which was provided in papers [15, 16].

All created networks were developed using Keras library with Tensorflow framework as a backend. Keras is an open source library written in Python. It is built on Tensorflow framework and contains numerous implementations of commonly used neural network building blocks and ready tools to preprocess images and text data. Keras offers a higher-level, more intuitive set of abstractions to develop deep learning models [17]. Moreover, this library allows to train networks on GPU.

TensorFlow is an open-source software library for high performance numerical computation. It is a symbolic math library, and is also used for machine learning applications such as neural networks. In this field Tensorflow aims at fast detection and classification of images, achieving the quality of human perception [18].

As shown in Fig. 5, U-Net network consists of two parts: an encoder (left) and a decoder (right). The encoder is a neural network which has a typical CNN architecture including four blocks. Each of these blocks consists of two convolutional layers with 3 × 3 filter, two ReLU activation functions and a maxpooling operation with 2 × 2 filter and step 2. The decoder contains the same number of blocks. Each decoder block consists of an upsampling operation with 2 × 2 filter merging with an appropriate map of features from the encoder, two convolutional layers with 3 × 3 filter and one ReLU activation function. The last layer of the network is convolutional with 1 × 1 filter, which classify every pixel. As a result, the network has 19 convolutional layers, 18 ReLU activation functions, 4 maxpooling operations, 4 upsampling operations, and 4 merge operations.

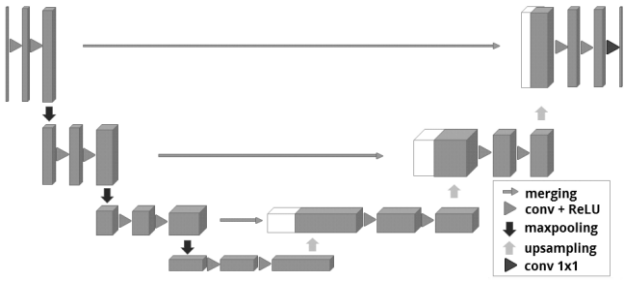

Fig. 5. – U-Net Architecture

As in the case of U-Net architecture LinkNet has the encoder and the decoder too. According to the network architecture, which is shown in Fig. 6, both subnets consist of 4 blocks. Each encoder block contains four convolutional layers, two merging operations and one maxpooling operation with 2 × 2 filter and step 2. Each decoder block has a similar architecture, except merging operations and a maxpooling operation, which was replaced with an upsampling operation with 2 × 2 filter. Before sending the map of features from an input data to the first encoder, there are implemented two operations of batch normalization, one ReLU activation function, one convolution with 2 × 2 filter and one maxpooling operation with 2 × 2 filter. After the last decoder there is performed a sequence of one upsampling operation with 2 × 2 filter, two operations of batch normalization, one ReLU activation function and one convolution with 2 × 2 filter.

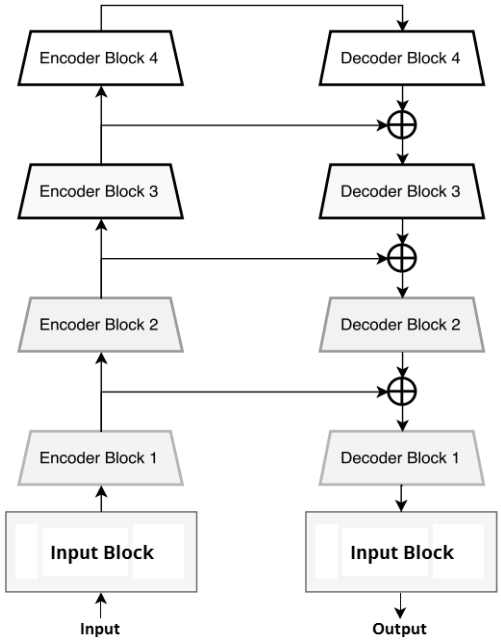

Fig. 6. – LinkNet Architecture

As the algorithm of numerical optimization Adam optimizer with a learning rate of 1e-3 was chosen. This optimizer combines best approaches from gradient descent and momentum optimizers and shows optimal and fast convergence for most tasks of machine learning [19]. Binary cross-entropy was chosen as the loss function. The batch consists of 18 samples. The training finishes after completing 96 epochs.

4. NUMERICAL RESULTS

Numerical experiments for developed algorithms were performed on satellite images of the private database. Information about the location of buildings was extracted from json files and generated as black-and-white masks, where a pixel is colored to white if it belongs to buildings.

The traditional approach for segmentation concerns the usage of parts of satellite images, which are fed to the input of CNN. Developed models require images of 512 × 512 pixels, so before the launch of deep learning algorithms each image and mask of dataset have been cut on parts of appropriate size by data windowing. Examples of sliced images and masks are shown in Fig. 7.

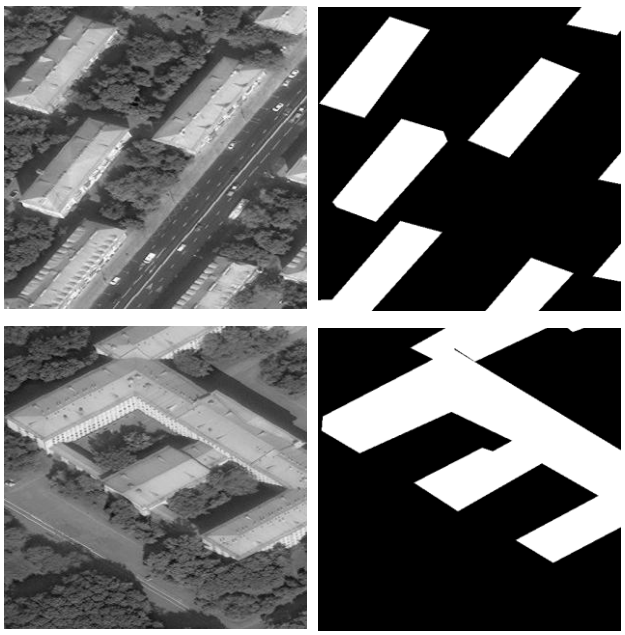

Fig. 7. - Examples of sliced images and masks

The prepared dataset contained 3264 images of 512 × 512 size. For numerical experiments it was shuffled and splitted on training and test set in ratio 80/20. In our research, only 2 classes were taken into account: “buildings” and “not-buildings”.

Developed CNNs were launched on NVIDIA DGX-1 supercomputer, which was provided by Artificial Intelligence Center of P.G Demidov Yaroslavl State University.

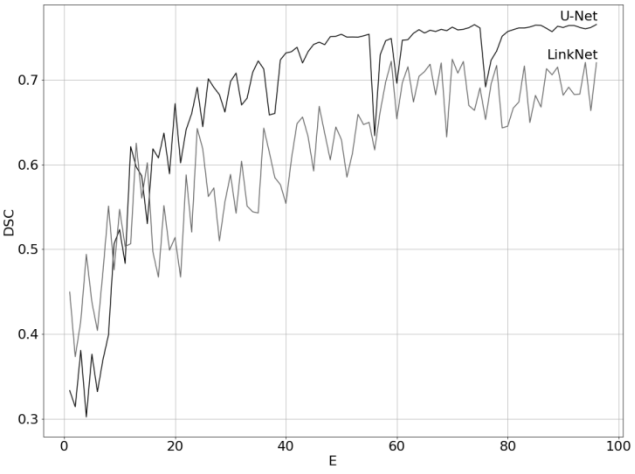

Fig. 8. - Dependencies of Sorencen-Dice coefficients on epochs for developed deep learning algorithms

As a rule, the quality of algorithms for image segmentation is evaluated by special coefficients for comparing the similarity of predicted and true masks. To estimate developed models there was used Sorensen-Dice coefficient (DSC). This index is binary measure of similarity, possesses the value from [0, 1] and can be calculated by the following formula:

$$DSC = \frac{2I}{S},$$

where $I = |X \cap Y|$ is a power of intersection and $S = |X| + |Y|$ is a sum of powers for real mask X and predictions Y. In our task, numerator I and denominator S can be calculated by following formulae

$$I = \sum_{\substack{x \in X \\ y \in Y}} xy, \qquad S = \sum_{\substack{x \in X \\ y \in Y}} (x + y),$$

where x, y are values of pixels from $[0, 1]$ for expert markup X and predicted masks Y respectively. Dependencies of DSC values from the number of epochs (E) for developed algorithms on the test subset are shown in Fig. 8.

Table 1. Numerical experiments

Model	Sorensen-Dice coefficient (DSC)
U-Net	0,77
LinkNet	0,72

According to the results presented in Fig. 8 and Table 1, the worst results of segmentation were shown by LinkNet, while the best results were obtained using U-Net. Some test results of U-Net are shown in Fig. 9.

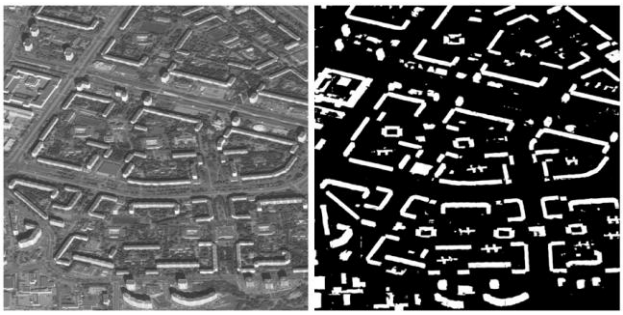

Fig. 9. – Test results

5. CONCLUSION

Numerical experiments of comparative analysis of developed algorithms were performed for the aerial photos of the private database. For modeling of numerical experiments there were extracted smaller images and masks, which were generated automatically from json files. Using the special metrics of similarity between expert markup and predicted masks there was shown that U-Net got better results compared with LinkNet. For U-Net the value of Sorensen-Dice coefficient (DSC) is equal to 0.77. All created networks are simple for its implementation.

The article was prepared with the financial support of the Ministry of Education of the Russian Federation as part of the research 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 also grateful to the AI-center of P.G. Demidov Yaroslavl State University for providing an access to NVIDIA DGX-1 supercomputer.

6. REFERENCES

[1] Y. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*. The MIT Press, 2016, 800 p.

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

[3] ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), Web: http://image-net.org/challenges/LSVRC/2012/.

[4] S. Seferbekov, V. Iglovikov, A. Buslaev, A. Shvets. Feature Pyramid Network for Multi-Class Land Segmentation. Web: https://arxiv.org/pdf/1806.03510.pdf.

[5] DeepGlobe. CVPR 2018 – Satellite Challenge, Web: http://deepglobe.org.

[6] 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. 234–341.

[7] G. Chhor, C. Bartolome Aramburu, I. Bougdal-Lambert. Satellite Image Segmentation for Building Detection using U-net. Web: http://cs229.stanford.edu/proj2017/final-reports/5243715.pdf.

[8] A. Chaurasia, E. Culurciello. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. Web: https://arxiv.org/pdf/1707.03718.pdf.

[9] The Cambridge-driving Labeled Video Database (CamVid), Web: http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/.

[10] IKONOS Satellite Images, Web: https://www.satimagingcorp.com/gallery/ikonos/.

[11] QuickBird 60cm Global High-Resolution Satellite Imagery, Web: http://www.landinfo.com/QuickBird.htm.

[12] DigitalGlobe satellite, Web: https://www.digitalglobe.com.

[13] DSTL Satellite Imagery Feature Detection, Web: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection.

[14] J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, 2016, 541 p.

[15] Khryashchev V.V., Priorov A.L., Pavlov V.A., Ostrovskaya A.A. Deep learning for region detection in high-resolution aerial images // Proceedings of 16-th IEEE East-West Design & Test Symposium (EWDTS'2016), Kazan, Russia, September 14 - 17, 2018. P. 792-796.

[16] V. Khryashchev, L. Ivanovsky, V. Pavlov, A. Ostrovskaya and A. Rubtsov, "Comparison of Different Convolutional Neural Network Architectures for Satellite Image Segmentation. *Proceedings of the 23rd Conference of Open Innovations Association FRUCT'23, Bologna, Italy.* 2018, pp. 172-179.

[17] A. Gulli, S. Pal. *Deep Learning with Keras*. Packt Publishing, 2017, 320 p.

[18] N. Shukla, *Machine Learning with Tensorflow*. Manning Publications, 2018, 272 p.

[19] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. Web: https://arxiv.org/abs/1412.6980.