Automatic detection of constructions using binary image segmentation algorithms

E A Dmitriev[1], A A Borodinov[1], A I Maksimov[1] and S A Rychazhkov[1]

[1]Samara National Research University, Moskovskoye shosse, 34, Samara, Russia, 443086

e-mail: dmitrievEgor94@yandex.ru, aaborodinov@yandex.ru

Abstract. This article presents binary segmentation algorithms for buildings automatic detection on aerial images. There were conducted experiments among deep neural networks to find the most effective model in sense of segmentation accuracy and training time. All experiments were conducted on Moscow region images that were got from open database. As the result the optimal model was found for buildings automatic detection.

1. Introduction

The automatically detecting objects in Earth remote sensing (RS) images task is one of the most difficult tasks. An example of a solution to the problem under consideration is [1]. Currently, one of the most effective approaches is semantic segmentation algorithms usage. In other words, for each image pixel, the object class to which it belongs is determined.

The segmentation of remote sensing images is used in many industries: geoinformatics, the creation of maps, analysis of land use, etc. At the moment, many segmentation process stages are solved manually with the help of operators, which leads to high economic costs in temporary resources, as well as some inaccuracies in the markup due to the human factor.

Currently, there are many algorithms for image segmentation [2, 3, 4], but the most effective are approaches using convolutional neural networks (CNN) [5]. For almost all computer vision tasks, convolutional networks provide more efficient results than other algorithms.

In recent years, various approaches have been proposed for the CNN models formation, which at the output give an original image segmentation map. One of the most effective methods is based on the use of fully connected neural networks [5]. Unlike the convolutional networks that are used for classification, there is no subnet of the multilayer perceptron for classification in fully connected networks.

The CNN architecture for semantic segmentation can be divided into two parts: the encoder and the decoder. The output coder produces feature maps with a smaller size than the input image. A decoder is used to restore the size of the feature maps. In the original versions of models of fully convolutional networks, the decoder was a geometric transformation to increase the size of images with various interpolation methods [5]. Currently, an approach is used where the decoder subnetwork is constructed symmetrically to the encoder's subnetwork with the exception of pooling layers. Instead of pooling layers, transposed layers [6] or unpooling layers [7] can be used.

The paper discusses 4 convolutional networks for detecting buildings with different encoder and decoder architectures. As the criteria for the algorithms effectiveness, network learning time and segmentation accuracy are used.

The work is organized in the following order. The second section describes the considered neural network architectures. The third section presents the experimental studies results on real images of the Moscow region. The final section summarizes the results and tells about the future research direction in the field of semantic segmentation algorithms.

2. Methods

As algorithms for binary semantic segmentation, we used SegNet neural networks [7], a model with an encoder from the ResNet-50 network [8] and a decoder in the form of a geometric transformation with bilinear interpolation, U-Net [9], LinkNet [6].

The SegNet network model is a classic encoder-decoder architecture. The SegNet encoder network consists of 13 convolutional layers which correspond to the first 13 convolutional layers in the VGG-16 network. The decoder architecture is almost symmetrical to the encoder's subnetwork, with the exception of pooling layers. In this paper, unpooling layers are used. The SegNet network model is shown in Figure 1.

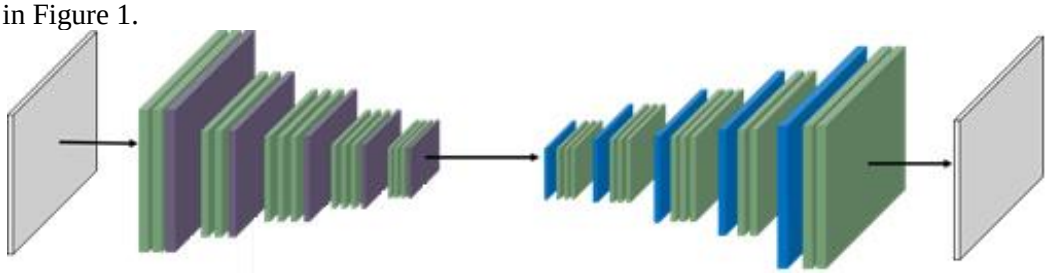

Figure 1. SegNet model.

The paper also considered a convolutional neural network for segmentation with an encoder based on ResNet-50. A feature of the ResNet-50 network is the use of residual connections, which make it possible to effectively solve the problem of a damped gradient arising with an increase in the number of neural network layers. The network model is shown in Figure 2.

The next neural network architecture under consideration is U-Net. The U-Net model feature is the feature maps concatenation on the lower and upper neural network levels. This approach is very similar to the residual connections in the ResNet-50 network, but in the case of U-Net, deeper connections are used. The network model is shown in Figure 3.

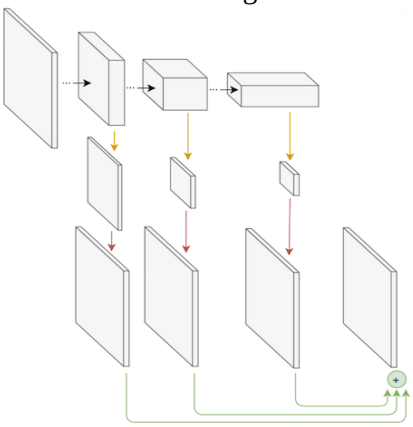

Figure 2. Fully convolutional network model based on ResNet-50.

LinkNet is an evolution of the U-Net model. The encoder and decoder are divided into several sub-blocks. LinkNet requires less computational resources in comparison with the considered models due to the rapid decrease in the size of attribute maps. At the network input, a decrease in feature maps occurs at the expense of pooling and convolution with a step equal to 2, and in the encoder block, at

the expense of convolution instead of pooling. In the decoder, transposed convolutional layers are used to restore the size of the images. The network model is shown in Figure 4.

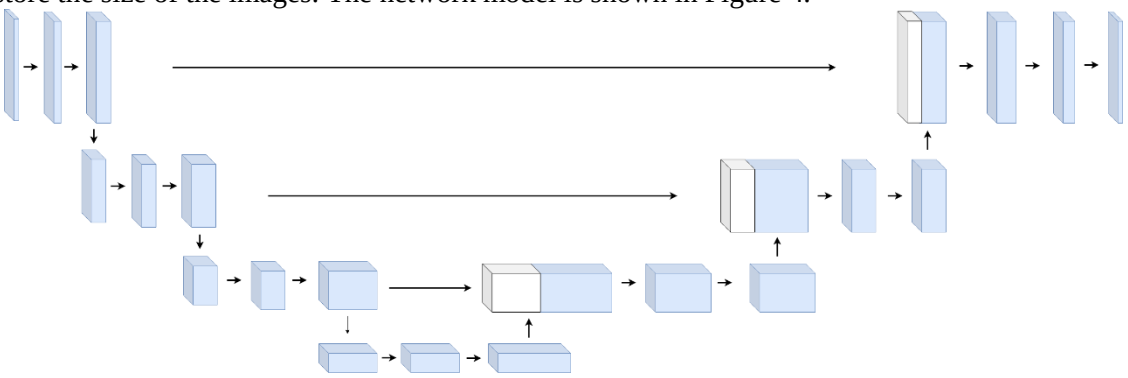

Figure 3. U-Net model.

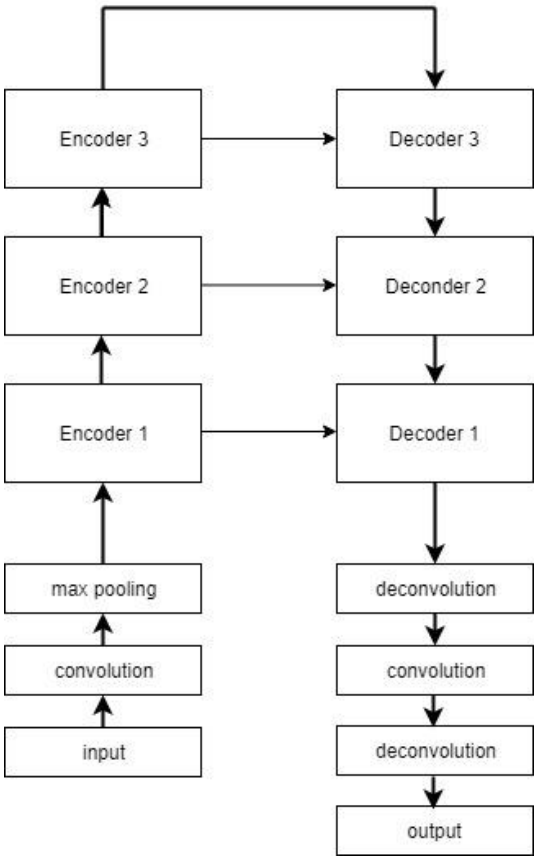

Figure 4. LinkNet model.

Cross-entropy was used as a loss function. According to [11], in the classification problem, the cross-entropy usage as a loss function allows achieving a better local minimum from the classification accuracy viewpoint with random algorithm parameters initialization compared to the standard deviation.

Let $I(n_1,n_2,n_3)$ – digital image applied to the input of the neural network, wherein $(n_1,n_2,n_3)\in\mathbf{D}$, $\mathbf{D}=\{(n_1,n_2,n_3): n_1=\overline{0,N_1-1}, n_2=\overline{0,N_2-1}, n_3=\overline{0,N_3-1}\}$, N_1,N_2 – the size of images, and N_3 – the number of channels in the input image. Let $Y(n_1,n_2,n_3)$ – mask the true segmentation, the dimensions of which coincide with the input image, and the number channels equal to the classes number. Each channel corresponded to a specific class. The classes were the buildings and the background. The values $Y(n_1,n_2,n_3)$ in the channels were 0 or 1, depending on the pixel class in the input image. Let

$O(n_1, n_2, n_3)$ – the image obtained at the neural network output whose size and the channels number coincide with the image markup. Let $y(n_3), o(n_3)$ – pixels with the same positions on the spaced and output images. Then the loss function as follows:

$$H(y,o) = -\sum_{i=0}^{N_3-1} y(i) \log o(i). \quad (1)$$

The target function performed functional mean error of the neural network training set. Let X^G – set with training images, where G – amount of elements, and w – neural network weights. Then the mean error is as follows:

$$Q(w, X^G) = \frac{1}{G} \sum_{i=0}^{G-1} \sum_{j=0}^{N_1-1} \sum_{k=0}^{N_2-1} H(O(i,j), Y(i,j)) \quad (2)$$

All models were trained using an adaptive stochastic gradient algorithm [12]. During the network training, the reducing technique the training coefficient was used in the event that the network quality value on the validation sample did not increase.

3. Experiments

The work considered photographs of settlements of the Moscow region [13]. RGB images of 512 × 512 size were fed to the network input. The number of shots was 3323. The ratio of the number of elements in the training sample to the number of elements of the test sample was 80:20. In the role of classes were the buildings and the background. An example of the image and mask is shown in Figure 5.

Figure 5. An example image and the mask part of the Moscow region settlement.

As can be seen from Figure 5, there were cases when the mask did not fully match the input image. Despite this, the inclusion of such images in the training sample made it possible to increase the metric value used in test images with an ideal mask even without a preprocessing stage.

The segmentation accuracy was used as a metric. The segmentation accuracy corresponds to the percentage of correctly classified pixels from the total number of pixels. The models were trained using the Nvidia GTX 1080 Ti graphics card. The experiments results are presented in table 1.

Table 1. The considered NN results.

Model	Training time, h	Segmentation accuracy, %
SegNet	4	96.7
Network based on ResNet 50	1.3	96.2
U-Net	2	96.9
LinkNet	**0.5**	**97.2**

According to the results of the experiments, it can be concluded that the approaches using transposed layers in the decoder used in LinkNet and the concatenation of the upper and lower feature maps used in the LinkNet and U-Net network allow to obtain higher generalizing abilities compared to other considered architectures. An example of the output image of the LinkNet network is shown in Figure 6.

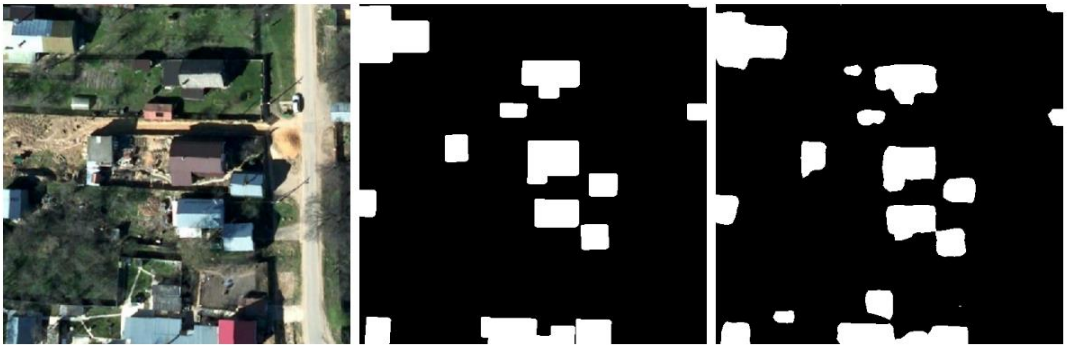

Figure 6. A test image example, markup and mask, obtained using LinkNet.

4. Conclusion

In this article, various convolutional neural networks architectures were investigated for the detection of structures in remote sensing images.

An experiments series was conducted, during which the optimal neural network architecture was identified in terms of training time and segmentation accuracy. Further research is planned on the use of conditional random fields to improve the segmentation quality.

5. References

[1] Myasnikov V V 2012 Method for detection of vehicles in digital aerial and space remote sensed images *Computer Optics* **36(3)** 429-438
[2] Kuznetsov A V and Myasnikov V V 2014 A comparison of algorithms for supervised classification using hyperspectral data *Computer Optics* **38(3)** 494-502
[3] Blokhinov Y, Gorbachev V A, Rakutin Y O and Nikitin A D 2018 A real-time semantic segmentation algorithm for aerial imagery *Computer Optics* **42(1)** 141-148 DOI: 10.18287/2412-6179-2018-42-1-141-148
[4] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* **20** 273-297
[5] Long J, Shelhamer E and Darrell T 2016 Fully convolutional networks for semantic segmentation *The Pattern Analysis and Machine Intelligence* **324** 100-108
[6] Chaurasia A and Culurciello E 2017 Linknet: Exploiting encoder representations for efficient semantic segmentation *IEEE Conference on Computer Visionand Pattern Recognition* **362** 234-247
[7] Badrinarayanan V, Kendall A and Cipolla R 2017 Segnet: A deep convolutional encoder-decoder architecture for image segmentation *IEEE Conferenceon Computer Vision and Pattern Recognition* **353** 125-145
[8] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *IEEE Conference on Computer Vision and Pattern Recognition* **123** 235-247
[9] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention – MICCAI* **345** 234-241
[10] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A C and Fei-Fei L 2015 ImageNet large scale visual recognition *IEEE Conference on Computer Vision and Pattern Recognition* **243** 121-136
[11] Golik P, Doetsch P and Ney H 2013 Cross-entropy vs. squared error training: a theoretical and experimental comparison *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 1756-1760
[12] Kingma D and Ba J 2014 Adam: A Method for Stochastic Optimization *International Conference on Learning Representations*
[13] Regional geographic information system of the Moscow region URL: https://rgis.mosreg.ru

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (RFBR) № 18-01-00748-a.