Augmentation Tools for Object Detection in Satellite Images by Using U-Net Neural Network

Ales Zhuk
Belarusian State University
Minsk, Belarus
ales.zhook@gmail.com

Sergey Ablameyko
Belarusian State University
United Institute of Informatics Problems
of NAS of Belarus
Minsk, Belarus
ablameyko@bsu.by

Abstract. **The paper proposes an approach to detect discrete objects on images, namely buildings using the U-NET neural network. The main idea of our approach is to use additional augmentations during model learning. The experiments carried out have shown good results.**

Keywords: **object detection, satellite images, neural network, U-Net, data set, data augmentations**

I. INTRODUCTION

Detecting and highlighting buildings on satellite images is an important task for various applications: building maps of the area, developing city infrastructure, searching for illegally built objects. Although the manual selection of buildings on satellite images is quite accurate, with a lot of images and the need for constant monitoring, manually processing them will take a lot of time and resources. Therefore, algorithms for automatic segmentation of satellite images are being developed. The task of automatic building detection can be complicated by bad weather conditions, the variety of shapes and colors of the found structures.

In recent years, neural networks have been used for image segmentation and processing. Classical neural network for object segmentation - U-NET. It was first used in 2015 for the segmentation of medical images [1]. The training set contained 30 images with 512x512 resolution. Dataset was expanded with additional transformations (rotations 90 degrees). The segmentation results surpassed other known methods and demonstrated the effectiveness of using U-NET on small image arrays.

Article [2] is devoted to the segmentation of satellite images. The task was to select 10 classes of objects (buildings / lakes / rivers / roads / etc.) on the images. The article describes the approach that was taken in the image segmentation competition on the Kaggle platform and helped the team to take third place. The idea was to use a modified U-NET network, and properties of some image channels (the images were 16-channel). So, water and vegetation could be detected without prior training, only by extracting information from image pixels. Due to the small number of images in the training set, data augmentation (rotations and flips) was applied. It should be noted that some images in the original training set are quite similar. For example, almost all buildings have blue roofs, making it easier to learn the network.

Typically, the U-NET is trained from scratch on some sort of initialized weights. The paper [3] demonstrates the possibility of using a pretrained network. And how U-NET can be improved using pretrained encoder. The neural network U-NET is described, in which VGG11, trained on the ImageNet weights [4], with a replaced fully connected layer was used as a contracting path (encoder). As a result of the work, the conclusions were next: the pretrained models converge faster to their limiting value, and that the recognition result of such a model is better in comparison to the non-pretrained network. Since this work was aimed at showing the benefits of using pre-trained networks, rather than getting the best result, there is still room for improvement. For example, using more complex networks as an encoder, such as VGG16, ResNet, etc.

We propose to introduce additional augmentations, such as adding noise, changing the brightness and contrast of the image, transforming perspective. And we show that this allows us to improve the segmentation result.

II. TRAINING SET

To solve the problem of segmentation of buildings, the set described in [5] was used. The images cover several settlements. The training set (similarly for the test set) contains 180 color three-channel images of 5000x5000 pixels with a spatial resolution of 0.3 meters. An example of an image and its mask is shown in Fig. 1. A mask is a binary image, where, depending on the pixel value (1 or 0), we determine whether this pixel belongs to the building or not.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

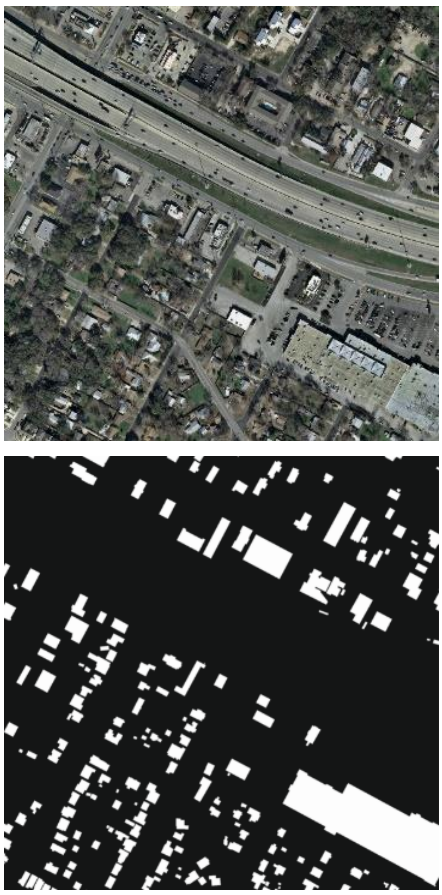

Fig. 1. Picture from training set

It should be noted that the images in the training set and the test set contain images of different cities. Training a neural network on images of some cities, and testing on others, allows you to understand how the algorithm adapts to other data of a similar nature.

III. Neural network architecture

In this work was used an architecture like the network described in [3]. U-NET consists of two parts: contracting and expanding. The contracting part is a convolutional network (convolution, activation, pooling), where the number of feature maps also increases on each layer. The expanding path is the opposite of the contracting path, where the pooling layer is replaced with an up-sampling layer, in which the image resolution is increased. U-NET also combines the features of the contracting path and with the expanding paths. The output of the U-NET network is a mask, where each pixel of the image is associated with the probability of its belonging to a particular class of objects. In our case, the probability that this pixel is a building. In our case, the contracting part was replaced by ResNet [6], pre-trained on the ImageNet weights.

IV. Neural network learning

Network learning parameters:

1. The original set (180 images) was divided into two: training (150) and validation (30). At each iteration, for each image of the training set, a 768x768 segment is randomly cut out, all such segments are grouped into batches and transmitted to the network input. The batch size was chosen 8 (the maximum possible with this image size and the provided graphics card).
2. Focal loss [8] was used as a cost function. If y_{ij} is a true value that determines the class of a particular pixel, $\overline{y_{ij}}$ is the probability of a pixel belonging to a class with label 1 obtained by the model. Let:

$$p_t = \begin{cases} \overline{y_{ij}} & , y_{ij} = 1, \\ 1 - \overline{y_{ij}} & , y_{ij} = 0. \end{cases} \quad (1)$$

Then the cost function can be written as:

$$C = -\frac{1}{n*m} \sum_{i=1}^{n} \sum_{j=1}^{m} -\alpha * (1 - p_t)^{\gamma} * \ln p_t \ , \quad (2)$$

where α, γ are parameters, which in our case are equal to 0.25 and 2, respectively. Focal loss can be characterized as a weighted cross-entropy function. Adding $\alpha * (1 - p_t)^{\gamma}$ to the cost function reduces the value of the function for well-classified objects and, consequently, improves the learning result for negative cases. The use of this function is necessary in the case of an unbalanced dataset when one of the classes prevails or concedes to the others. In our case, buildings occupy 0.15 of the images in the training set. And the use of focal loss avoids preprocessing associated with building a balanced training set.

3. At the stage of network learning image augmentations (described in the next section) were additionally performed.
4. We used the Adam optimization algorithm [9] with a learning factor of 0.0001. (During the training of the model, the learning factor was decreased several times. The criterion for the decrease is that the validation metrics stop improving or changing.)

V. Augmentations

Standard data augmentations used in satellite imagery segmentation tasks are rotations by angles divisible by 90 degrees and image flips. In this work, we propose and apply additional image transformations. The final set of augmentation is listed below:

- Rotate at a random angle multiple of 90 degrees,

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

- Vertical flip,
- Horizontal flip,
- Adding Gauss noise,
- Change hsv. (Random change in the hue, saturation, and brightness of the color of an image within a certain range),
- Change the brightness and contrast of the image (Random change in the brightness and contrast of the image in a certain range),
- Transformation of the perspective of the image.

At each training iteration, for a particular image, the probability that some augmentation would be applied was 0.25.

To test and implement augmentations, we used the library described in [10].

VI. Experimental results

In our work, we took advantage of the cloud computing capabilities provided by Google Colaboratory [12]. We were allocated a Tesla P100 graphics card with 16 GB of memory. This allowed us to use a more complex architecture of the neural network, increase the batch and size of images at the input of the neural network, carry out experiments faster and do more iterations during training. The result was assessed using next metrics: accuracy (3) and Jaccard coefficient (4)

$$A = \frac{1}{n*m}\sum_{i,j}^{n,m}\begin{cases}1, \text{где } y_{ij} = \bar{y}_{ij},\\ 0\end{cases} \quad (3)$$

$$J = \frac{1}{n*m}\sum_{i,j=1}^{n,m}\frac{y_{ij}*\bar{y}_{ij}}{y_{ij}+\bar{y}_{ij}-y_{ij}*\bar{y}_{ij}}, \quad (4)$$

where y_{ij} is the true pixel value, $\overline{y_{ij}}$ is the model predicted value, $n * m$ is the image size. The TABLE 1 below shows a comparison of the results of several experiments: using standard augmentation, additional augmentation and solution [11]. This metrics show that the best building segmentation result was obtained when we used additional image augmentations during network learning.

VII. Discussion and conclusion

A few examples of building segmentation in Fig. 2.

It should be noted that to improve the result of the selection of discrete objects in satellite images, various transformations of images should be used. In our case, changing the brightness and saturation of colors, adding noise, changing the perspective of the image helped to increase the resulting metric (Jaccard coefficient) from 74.12 to 75.78.

The use of other neural networks can be considered as further improvements. For example, instead of the pre-trained ResNet34 network, take DenseNet or SE-ResNet as a basis. Try to predict the result using an ensemble of several networks, that is, determine the class of each pixel not according to the output of one of the networks, but based on a certain rule and the output of several networks at once. And to practice more with image augmentations.

Fig. 2. Some recognition results

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

TABLE 1. Results Comparison Table

Model described above with standard augmentations						
Metrics	***Bellingham***	***Bloomington***	***Innsbruck***	***San Francisco***	***Tyrol***	***Overall***
Jaccard index	69.03	73.44	74.50	75.02	76.54	**74.12**
Accuracy	96.89	97.45	96.88	91.31	97.87	96.08
Model described above with additional augmentations						
Jaccard index	69.95	75.19	75.46	77.29	77.69	**75.78**
Accuracy	96.96	97.61	97.06	92.26	98.00	96.38
Solution [11]						
Jaccard index	69.75	72.04	74.64	74.55	77.40	**73.91**
Accuracy	96.77	97.13	96.83	91.14	97.92	95.96

REFERENCES

[1] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI, Springer, 2015, pp. 234–241.

[2] V. Iglovikov, S. Mushinskiy, V. Osin, Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition, arXiv preprint, 2017, 6 p.

[3] V. Iglovikov, A. Shvets, TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation, arXiv preprint, 2018, 5 p.

[4] O. Russakovsky, J. Deng, H. Su, et al., ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision – Springer, 2015, pp. 211–252.

[5] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark, IEEE International Symposium on Geoscience and Remote Sensing Symposium (IGARSS), July 23-28, IEEE, 2017, pp. 3226–3229.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for

[7] Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 27-30, IEEE, 2016, pp. 770–778.

[8] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, Focal Loss for Dense Object Detection, IEEE International Conference on Computer Vision (ICCV), Venice, 22-29 Oct., IEEE, 2017, pp. 2999–3007.

[9] D. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations, Banff, Canada, April 14-16, 2014, 15 p.

[10] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, Albumentations: Fast and flexible image augmentations, Information, MDPI, 2020, vol. 11, 4 p.

[11] Girard, N., Polygonal Building Segmentation by Frame Field Learning, Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka, arXiv preprint, 2020, 30 p.

[12] Google Colaboratory [Electronic resource]

[13] https://colab.research.google.com. Date of access: 06.09.2020

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.