Deep learning for region detection in high-resolution aerial images

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

Andrey Priorov
P.G. Demidov Yaroslavl State University, Russia
andcat@yandex.ru

Vladimir A. Pavlov
P.G. Demidov Yaroslavl State University, Russia
vladimir@1pavlov.com

Anna A. Ostrovskaya
People's Friendship University of Russia, Russia
ostrovskaya_aa@rudn.university,

Abstract

The goal of given investigation is to develop deep learning and convolutional neural network methods for automatically extracting the locations of objects such as water resource, forest and urban areas from given aerial images. We show how deep neural networks implemented on modern GPUs can be used to efficiently learn highly discriminative image features. For deep learning on supercomputer NVIDIA DGX-1 we used the marked image database UrbanAtlas, which contains images of 21 classes. Images obtained from the Landsat-8 satellites are used for estimation of automatic object detection quality. Object detection on aerial images has found application at urban planning, forest management, climate modelling, etc.

1. Introduction

The convolutional neural network (CNN) algorithm used in this study is based on the combination of deep learning algorithms and advanced GPU technology [1]. The main advantage of CNN algorithms 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 [2-6]. Deep learning implements a neural network approach to "teach" machines object detection and classification [7]. While neural network algorithms have been known for many decades, only recent advances in parallel computing hardware have made real-time parallel processing possible [8, 9]. Essentially, the underlying mathematical structure of neural networks is inherently parallel, and perfectly fits the architecture of a graphical processing unit (GPU), which consists of thousands of cores designed to handle multiple tasks simultaneously [10-14]. The software's architecture takes advantage of this parallelism to drastically reduce computation time while significantly increasing the accuracy of detection and classification.

As in other problems solved with the help deep learning methods, the existence of a large database of annotated images is the most important factor. Let's consider this problem in more detail. There are several publicly available datasets of annotated images from satellite imagery.

DeepSat. DeepSat [15] was published in 2015. It contains two sets of annotated images from different satellite constellations: 500,000 Sat-4 images, divided into 4 land-use classes ("barren land", "trees", "grassland" and a class that consists of all land cover classes other than the above three), 405 000 Sat-6 images, divided into more than 6 classes of land use ("barren land", "trees", "grassland", "roads", "buildings" and "water bodies"). All samples have a size of 28 × 28 px at a spatial resolution of 1 m / px and contain 4 channels (red, green, blue and NIR - near infrared radiation). It was shown in [16] that it is possible to create a classifier based on convolutional networks with a classification accuracy of about 99% using such set of images. However, while this dataset is very useful for preliminary preparation of more complex models (for example, image segmentation), it does not allow to take further steps for detailed analysis of land-use and comparison of urbanized environments.

UCMerced. This Dataset of annotated images was published in 2010 [17] and contains 2100 images of 21 land-use classes with a size of 256 × 256 pixels, with a resolution of 1 m / px. This dataset is considered to be

978-1-5386-5710-2/18/$31.00 ©2018 IEEE

a "solved problem", since a modern neural network based on classes [18] achieves> 95% accuracy.

UrbanAtlas. The European Urban Atlas [19] provides reliable, inter-comparable, high-resolution (2.5 meters per pixel) 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. Images obtained from the Landsat-8 satellites [20] are used for estimation of automatic object detection quality. Examples of images from Landsat-8 satellites are shown on Fig. 1. Landsat 8 images have a resolution of 30 meters per pixel. This is the highest resolution from open sources aerial images.

The second part of this article describes the principle of optimizing a U-NET neural network for the analysis of space images. The third part describes the results of the experiment on the detection of 3 classes of objects: water, forest, agriculture. Fourth and fifth parts contain the main conclusions on this work.

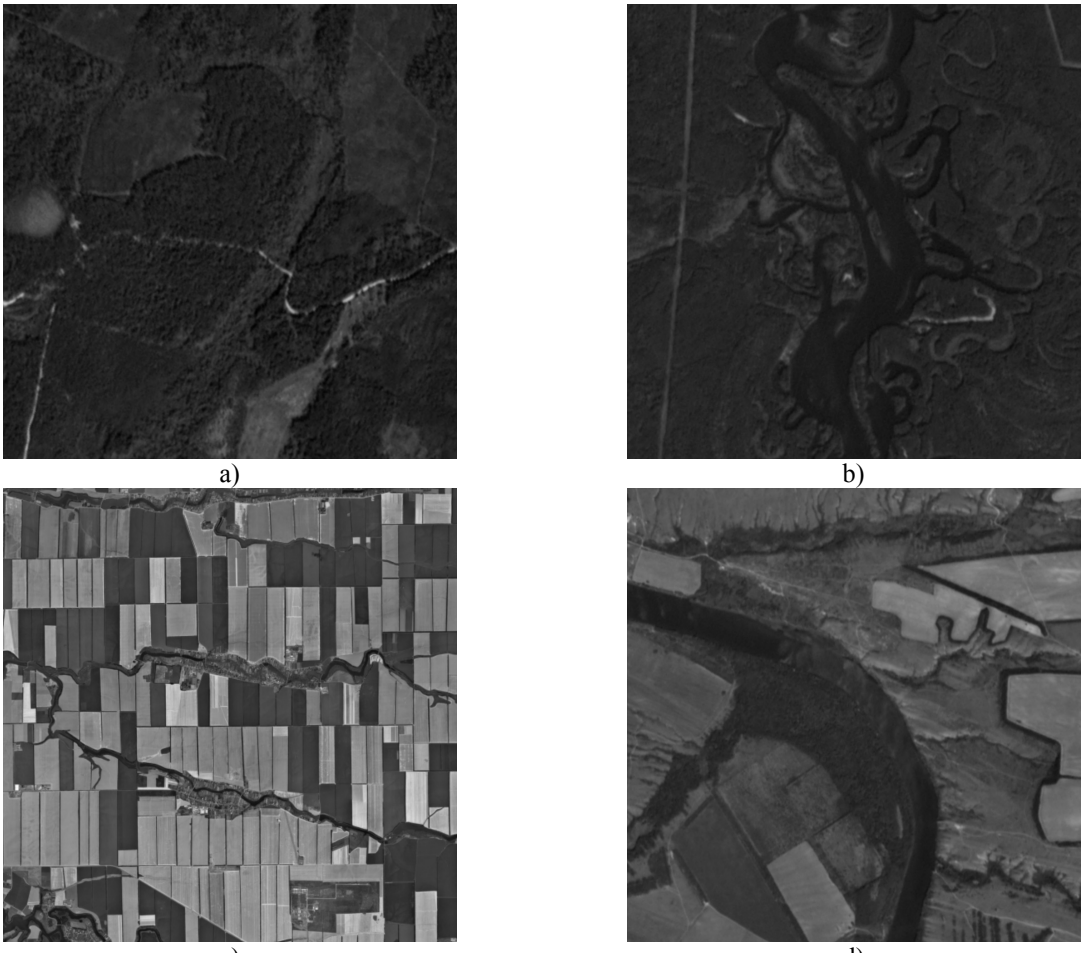

Figure 1. Examples of images from Landsat-8:
a) scene with class "forest", b) scene with classes "forest" and “water resource”,
c) scene with class "agriculture", d) scene with classes "forest",
“agriculture” and “water resource”

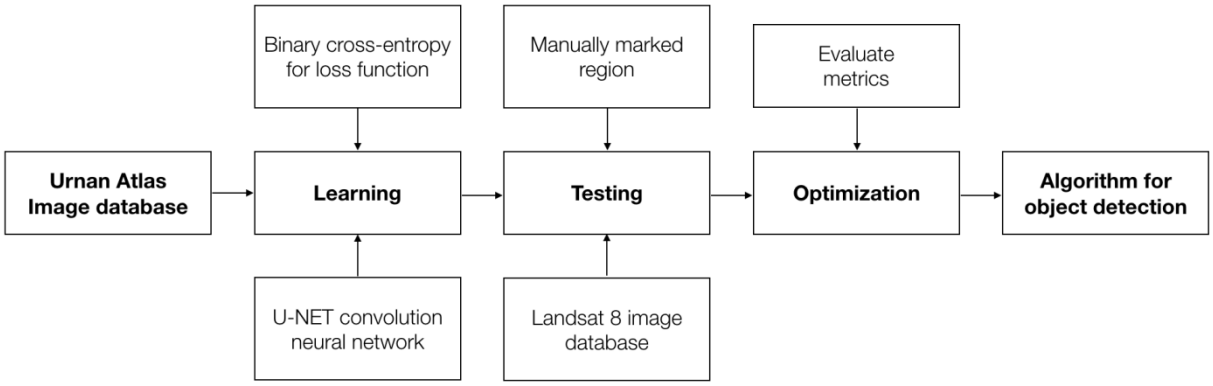

Figure 2. Scheme of learning, testing and optimization of convolutional neural network

2. CNN learning, testing and optimization

The scheme of the research is shown in Fig. 2. It contains 3 key steps: training, testing and optimization. Performing these steps leads to the construction of the final algorithm for detecting objects on images from Landsat-8.

In this research we used the U-NET architecture of a convolutional neural network presented on Figure 4. The network consists of two parts: a scraping conveyor (on the left) and an expanding network (on the right), which are represented by 23 convolutional layers [21-22].

The testing was carried out on the supercomputer NVIDIA DGX-1 in the Center for Artificial Intelligence of the Yaroslavl University named by. P.G. Demidov. The total network training time on the supercomputer is 5 hours. Network training can be done at any GPU or CPU, but the learning time will increase tens of times. Increasing the number of detection classes will increase the learning time linearly if we supply the same amount of new training data without increasing the number of epochs.

An error function is used that is minimized in the controlled learning of the neural network to evaluate the accuracy of the learning and testing stages of the neural network. For classification problems usually use category-based cross-entropy like an error function

$$H = \sum_{i=1}^{N} y_i' log(y_i), \quad (1)$$

where y_i' – true probability distribution or i-th class, y_i – distribution of probability of detection for i-th class, N – number of classes (in our research we have 3 classes: "Forest", "Agriculture" and "Water".).

In our problem of detection, classes are not mutually exclusive and to use of formula (1) is not appropriate. In our experiment, we applied a compound loss function that includes binary cross-entropy and an approximate Jacquard distance:

$$loss = bce - log(ja), \quad (2)$$

where bce – binary cross-entropy:

$$bce = -\sum_{i=1}^{N}(y_i' \, log \, y_i + (1 - y_i') \, log(1 - y_i)). \quad (3)$$

The approximate Jacquard distance which calculated by the following formula:

$$ja = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum y_i' * y_i}{\sum y_i + \sum y_i - \sum y_i * y_i'}. \quad (4)$$

The above expressions allow us to estimate the measure of similarity between N classes (3 in our case).

3. Experimental results

We created an experiment to score the quality of our detectors. We compare the results from new detectors and expert marking. The test dataset contains100 space images. Each picture was marked by 10 experts to exclude the subjective factor, and the final polygon is the average contour of expert marking. The accuracy of the marking is the percentage of the intersection of the area allocated by the detector and the expert multi polygon

$$R = \frac{S_{pred} \cap S_{exp}}{S_{exp}}. \quad (5)$$

Where S_{pred} – detected multipolygon, S_{exp} – expert multipolygon. The results of the experiment are shown

in Fig. 3. Each chart has two parts. The light part is the detector of the first version, the dark one is the increase in the detection accuracy of the detector of the second version with Jacquard distance.

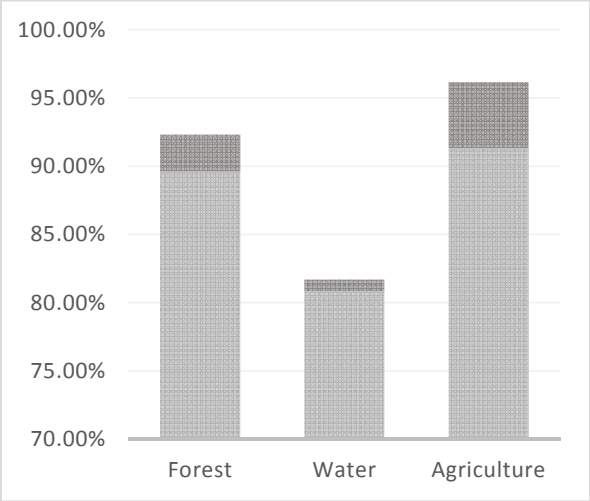

Figure 3. The diagram of accuracy of detection of 3 classes: forest, agriculture and water

As seen from the experiment, the choice of another loss function gave a significant increase in the detection accuracy. Detection of each class has improved by at least 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.

4. Conclusion

The developed algorithms are based on the implementation of a relatively new approach in the field of satellite image analysis – deep learning and a convolutional neural network. We show how deep neural networks implemented on modern GPUs can be used to efficiently learn highly discriminative image features. 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.

5. Acknowledgments

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.

6. References

[1] T. Qu, Q. Zhang, S. Sun, “Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks”, Multimedia Tools and Applications, Volume 76, Issue 20, pp. 21651–21663, October 2017.

[2] E. P. Baltsavias, “Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems”, ISPRS Journal of Photogrammetry and Remote Sensing, 58(3-4): pp.129-151, January, 2004.

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

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

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

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

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

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, “Overfeat: integrated recognition, localization and detection using convolutional networks”, International Conference on Learning Representations, 2013.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, “DeCAF: a deep convolutional activation feature for generic visual recognition”, ICML’14,

Proceedings of the 31st International Conference on Machine Learning, volume 32, pp. I-647-I-655, Beijing, China 2014.
[10] H. Mayer, “Object extraction in photogrammetric computer vision”, ISPRS Journal of Photogrammetry and Remote Sensing, 63(2):213-222, March 2008.
[11] O. A. B. Penai, K. Nogueira, J. A. dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?”, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 44–51, 2015.
[12] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, Web: https://arxiv.org/abs/1409.1556, 2014.
[13] M. Zhai, Z. Bessinger, S. Workman, N. Jacobs, “Predicting ground-level scene layout from aerial imager”, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, pp. 867-875, 2017.
[14] N. Jean, M. Burke, M. Xie, W. Davis, D. Lobell, S. Ermon, “Combining satellite imagery and machine learning to predict poverty”, Science 353, 6301, pp. 790–794, 2016.
[15] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, R. Nemani, “DeepSat - A learning framework for Satellite Imagery”, Proceedings of SIGSPATIAL’15, Bellevue, WA, USA, 2015.
[16] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko, K. Karantzalos, “Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data”, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 83–88, June 2016.
[17] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition”, In 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770–778, 2016.
[18] M. Castelluccio, G. Poggi, C. Sansone, L. Verdoliva, “Land use classification in remote sensing images by convolutional neural networks”, Web: https://arxiv.org/abs/1508.00092, 2015.
[19] “European Union. 2011. Urban Atlas”, Web: https://www.eea.europa.eu/data-and-maps/data/urban-atlas.
[20] “Landsat8”, Web: https://landsat.usgs.gov/landsat-8.
[21] 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: pp. 234-241, 2015.
[22] V. Khryashchev, V. Pavlov, A. Priorov, E. Kazina, “Convolutional Neural Network for Satellite Imagery”, Proceedings of the 22th Conference of Open Innovations Association FRUCT'22. Jyvaskyla, Finland, pp. 344-347, 2018.