Detection and recognition of objects on aerial photographs using convolutional neural networks

V A Pavlov and M A Galeeva
Peter the Great St. Petersburg Polytechnic University, Polytehnicheskaya ul., 29, St. Petersburg, Russia
E-mail: pavlov_va@spbstu.ru

Abstract. The article is devoted to the use of neural network methods to solve the problem of detection and classification of small objects on aerial photographs. The architectures of YOLO 2 and YOLO 3 are discussed. The training procedure is described, the obtained results are analyzed. During the work it was shown that the considered architectures can be used to solve the problem of detecting and classifying objects in images obtained as a result of aerial photography.

1. Introduction
Intelligent Systems of Computer Vision are actively used in different fields: industry [1, 2], space [3–5], visual tracking [6–8], medicine [9, 31–36], road safety management automation [30], etc. The main idea of the computer vision system is image processing. The computer vision system can solve problems of detection, recognition, creation of new scenes and images from existing ones (transfer of styles) by using information from a video stream obtained by a camera. The actual task is to detect and classify small-sized objects in images obtained as a result of aerial photography using an aircraft or a space satellite. For the case of camera location on board of an aircraft the solution of the problem is complicated by different factors. These factors are: the presence of different types of objects, blurring [10], variations in lighting, small sizes of objects, compression artifacts, background noise, as well as geometric transformation of the image.

Two types of methods are used to solve the problem of objects detection and recognition [11]: two-stage [12–17] and one-stage [18–21]. The second type of methods has a higher speed and is not inferior to the methods of the first type. A review of the literature [21–22] showed that the approaches of the YOLO family [18, 20, 21] are the most effective at the moment. For further research, architectures YOLO 2 and YOLO 3 were chosen. Unlike the first version YOLO, fully connected layers are removed, and the input resolution of the network may change during the detection process.

The purpose of this work is to survey the accuracy of the detection and recognition algorithms YOLO 2 and YOLO 3, trained on images obtained as a result of aerial photography obtained from an aircraft.

2. Architectures description
Consider the architecture and principle of operation of YOLO 2 and YOLO 3. To extract features from images of objects in YOLO 2, the Darknet-19 convolutional neural network model is used. [20]. The network consists of 19 convolutional layers, 5 max pooling layers, global average pooling layer and softmax. Darknet-19 mainly uses $3 \times 3$ convolution filters. Instead of fully connected layers, a global
average pooling layer is applied in front of the softmax layer in Darknet-19. Convolutional neural network, containing this layer in the structure is fully convolution [24]. As shown in [20], it can be applied to images of different resolutions without changing parameters.

Darknet-19 and YOLO 2 use batch normalization for regularization, convergence and stabilization of training.

In [20] Darknet-19 trained on a set of 160 epochs with ImageNet 1000 data. During training, the authors used standard methods of artificial expansion of the set of training examples. After training on images with a resolution of $224 \times 224$, the authors increased the resolution to $448 \times 448$ and trained the network for another 10 epochs.

To solve the detection problem, the authors in [20] modified Darknet-19 by removing the last convolutional layer. Three convolutional layers with 1024 filters with a convolution size of $3 \times 3$ each are added. They are followed by the last convolutional layer with a convolution size of $1 \times 1$ and the number of outputs that is necessary for the detection and classification of $N_C$ classes.

Since the convolutional layers gradually reduce the size of feature maps, there are difficulties in finding objects that are small in size relative to the size of the image. Therefore, feature maps from output of the 13th convolutional layer are combined with feature maps from layer 20 and passed to the input of layer 21, so that the model can use fine grained features.

The entire input image, which is passed to the input of the network, is divided into a grid of $S \times S$ cells. Instead of predicting the position of the bounding boxes, as is done in YOLO, in YOLO 2 the displacement of anchor boxes in each of the $S \times S$ cells is predicted. Anchor boxes are formed before the start of training based on the training data set using the $k$-means method. For each anchor box, the network predicts: 4 values $(x, y, w, h)$ for correcting the size and position of the anchor box, confidence score and $N_C$ probabilities by class. $(N_C + 5) \times B$ values are formed for a single cell, where $B$ – number of anchor boxes (in [20] $B=5$).

During YOLO 2 training, after every 10 batches, the resolution of the input images may change in a random order. This feature allows to detect objects with different sizes. The size can be changed in increments that are multiples of 32.

The softmax function is used to calculate probabilities for class in YOLO 2. The sum of all probabilities is 1. Thus class labels are mutually exclusive. Softmax is used in assumption that there is only one class in every bounding box. In YOLO 3 for each bounding box several classes are predicted with using an independent logistic classifiers. This is how multilabel classification is realized. The method allows better modelling of processed data. Binary cross-entropy is used for each label to calculate the losses. This approach reduces the computational load.

In YOLO 3 the presence of an object for each bounding is box predicted by using logistic regression. If one of the bounding (anchor) boxes overlaps with ground-truth box more than others, the objectness value is 1 in this case. If the overlapping of other anchor boxes is more than some threshold value (which is 0.5 by default), these predictions are not used.

In contrast, in YOLO 3 the prediction is performed for three different scales according to the principle similar to the FPN (feature pyramid network) [23] work. Different scales are set by lowering the dimension of the sampling of the input image by 32, 16 and 8, respectively. After the predictions for two scales, the feature maps are upsampled by 2 times. Then concatenation with earlier maps of features of the same dimension is performed. This allows both low-level and higher-order features to be used together. The results for the three scales are processed by the non-maximum suppression algorithm.

Anchor boxes as well as in YOLO 2, formed by a clustering algorithm $k$-means. A total of 9 rectangles are used, divided into three groups depending on the scale. For feature extraction and classification in YOLO 3, Darknet-53 is used, which contains 53 convolutional layers. It consists of $3 \times 3$ and $1 \times 1$ layers and residual blocks, which allow adding the previous output values to the output of the current layer. This approach allows to solve the problem of vanishing gradient of deep neural network. In YOLO 3 instead of pooling layers convolution with stride 2 is used.
The considered architectures YOLO 2 and YOLO 3 show high results in comparison with other modern algorithms for the detection and classification of objects. The value of mAP for YOLO 2 is 78.6%, which is the biggest in comparison with YOLO, Faster R-CNN, and others [20]. YOLO 3 has even bigger mAP (in comparison to SSD, RetinaNet and others). The processing speed of YOLO 3 is also significantly higher [21].

3. Results

The results of the work were obtained using computational resources of Peter the Great Saint-Petersburg Polytechnic University Supercomputing Center (www.scc.spbstu.ru). The training was done on two NVIDIA Tesla K40X GPU. The training process lasted 200,000 iterations. To create the dataset the following sources were used: aerial photographs from the aircraft, «DOTA: A Large-scale Dataset for Object DeTection in Aerial Images» [25]; «VIVID Dataset» [26]; kaggle service [27]. Google maps [28]. The training dataset contains 7869 images (87287 objects), validation dataset contains 2082 images (22094 objects) and test dataset contains 1504 images (15800 objects). Input resolution 608×608. The dataset contains 6 classes: airplane, building, car, helicopter, ship, large vehicle. The large vehicle class includes buses and trucks.

Metrics based on the calculation of the Intersection over Union (IoU) [29] were used to determine the accuracy of detection and classification. The IoU metric is a ratio of two areas of rectangles: intersection area of the detected rectangle and ground truth-rectangle to the area of their union. Precision and recall values are calculated based on IoU. Precision shows the percentage of true predictions from the number of all detections that the system considered positive, recall – the percentage of positive detections from all true objects. To evaluate the classification accuracy for each class of objects used metric AP (average precision) [29]. AP is calculated as an average precision for 11 recall levels. The sum of the AP values, averaged over number of classes, is a metric mAP (mean average precision) [29].

Table 1 shows the results of detection and classification: AP values for all classes and averaged mAP and IoU. Both considered architectures show good results: values AP and mAP exceed 50%. From this table it is clear that the accuracy of YOLO 3 is higher compared to YOLO 2. The largest performance gain YOLO 3 shows for a car class (more than 20%). The average IoU value for YOLO 3 is 10% higher.

| Parameter                        | YOLO 2   | YOLO 3  |
|----------------------------------|----------|---------|
| Average precision (AP) for airplane | 88.09    | 88.66   |
| Average precision (AP) for building | 70.81    | 72.33   |
| Average precision (AP) for car   | 65.62    | 86.49   |
| Average precision (AP) for helicopter | 82.81    | 87.62   |
| Average precision (AP) for ship  | 84.97    | 86.37   |
| Average precision (AP) for large vehicle | 65.29    | 76.50   |
| mean AP (mAP)                    | 76.14    | 82.99   |
| Jaccard index or Intersection over Union (IoU) | 50.71    | 61.37   |

Figures 1-2 show examples of YOLO 2 and YOLO 3. Both architectures handle the detection of objects of different sizes. The result of detection of YOLO 2 is not accurate enough in the case of the presence in the frame of a plurality of objects of small size close to each other. YOLO 3 was able to detect and recognize much more objects on the frame.
Figure 1. The results of YOLO 2.

Figure 2. The results of YOLO 3.
4. Conclusions
In this paper we have considered the use of the objects detector and classifier based on the neural network to solve the problem of detection and classification of objects in the images obtained as a result of aerial photography. For the study were chosen modern architectures YOLO 2 and YOLO 3. The training is performed on images from different datasets, including our own. The results of experiments on the test set of images showed that both architectures achieved high results. YOLO 3 is superior to YOLO 2 by the values of AP, mAP, and the accuracy of objects localization. YOLO 3 is better at detecting densely concentrated objects of small size. The detector based on YOLO 2 and YOLO 3 can be used in the tasks of detecting small objects on aerial photographs.

References
[1] Bakir N, Pavlov V, Zavjalov S, Volvenko S, Gumenyuk A and Rethmeier M 2019 Development of a novel optical measurement technique to investigate the hot cracking susceptibility during laser beam welding Welding in the World 63 435–441.
[2] Bakir N, Pavlov V, Zavjalov S, Volvenko S, Gumenyuk A and Rethmeier M 2018 Novel metrology to determine the critical strain conditions required for solidification cracking during laser welding of thin sheets Journal of Physics: Conference Series 1109.
[3] Fomin I, Orlova S, Gromoshinskii D and Bakhshiev A 2018 Object detection on docking images with deep convolutional network Studies in Computational Intelligence 799 136-143
[4] Bobrovsky A, Galeeva M, Morozov A, Pavlov V and Tsytuslin A 2019 Automatic detection of objects on star sky images by using the convolutional neural network Journal of Physics: Conference Series 1236 1066
[5] Devyatkin A, Tsytuslin A, Bobrovsky A, Morozov A, Gorshanov D and Pavlov V 2019 Adaptation of frame frequency to observation stages at control of spacecraft convergence Journal of Physics: Conference Series 1236 1069
[6] Zanina M, Pavlov V, Zavjalov S and Volvenko S 2018 TLD Object Tracking Algorithm Improved with Particle Filter 41st International Conference on Telecommunications and Signal Processing (TSP) 1-4.
[7] Pavlov V, Zavjalov S, Volvenko S and Zanina M 2018 Modification of Tracking Algorithm Struck by the Application of the Scale Estimation Method 41st International Conference on Telecommunications and Signal Processing (TSP) 1-5
[8] Abramov N, Bolsunovskaya M, Lekashow A and Barinov D 2016 Algorithms for detecting and tracking of objects with optical markers in 3D space Proceedings of the 19th International Conference on Soft Computing and Measurements, SCM 362-364
[9] Abramov N, Fradkin M, Rouet L and Wischmann H 2017 Configurable real-time motion estimation for medical imaging: application to X-ray and ultrasound Journal of Real-Time Image Processing 13 147-160
[10] Kobobynkov A, Grishentsev A, Velichko E, Korikov C, Aleksanin S, Fedosovskii M and Bondarenko I 2016 Calculation of regularization parameter in the problem of blur removal in digital image Optical Memory and Neural Networks (Information Optics) 3 184-191
[11] Haykin S 1998 Neural Networks: A Comprehensive Foundation (Prentice Hall) 842
[12] Dalal N, Triggs B 2005 Histograms of oriented gradients for human detection Computer Vision and Pattern Recognition (CVPR) 1 886–893
[13] Viola P, Jones M J 2001 Rapid Object Detection using a Boosted Cascade of Simple Features proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 511
[14] Felzenszwalb P F, Girshick R B, McAllester D and Ramanan D 2010 Object detection with discriminatively trained part based models IEEE Transactions on Pattern Analysis and Machine Intelligence 32 1627-1645
[15] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation Computer Vision and Pattern Recognition (CVPR) 580-587
[16] Girshick R 2015 Fast R-CNN Proceedings of the International Conference on Computer Vision (ICCV) 1440-1448

[17] Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: Towards real-time object detection with region proposal networks Neural Information Processing Systems (NIPS)

[18] Redmon J, Divvala S, Girshick R and Farhadi 2016 A You Only Look Once: Unified, Real-Time Object Detection (Preprint arXiv:1506.02640v5)

[19] Liu W, Anguelov D, Erhan D, Szegedy C and Reed S E 2016 SSD: single shot multibox detector European Conference on Computer Vision (ECCV) 21-37

[20] Redmon J and Farhadi A 2017 YOLO9000: Better, Faster, Stronger IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 7263-7271

[21] Redmon J, Farhadi A 2018 Yolov3: An incremental improvement (Preprint arXiv:1804.02767)

[22] Lin T et al 2017 Focal loss for dense object detection (Preprint arXiv arXiv:1708.02002).

[23] Lin T, Dollár P, Girshick R, He K, Hariharan B and Belongie S 2017 Feature Pyramid Networks for Object Detection (Preprint arXiv:1612.03144)

[24] Lin M, Chen Q and Yan S 2013 Network In Network (Preprint arXiv: arXiv:1312.4400v3)

[25] URL: captain-whu.github.io/DOTA/index.html

[26] URL: vision.cse.psu.edu/data/vividEval/datasets/datasets.html

[27] URL: kaggle.com

[28] URL: google.ru/maps

[29] Everingham M, Van Gool L, Williams C, Winn J and Zisserman A 2010 The PASCAL Visual Object Classes (VOC) Challenge Int J Comput Vis 88 303-338

[30] Makarov A, Bolsunovskaya M and Zhigunova O 2018 Comparative analysis of methods for keypoint detection in images with different illumination level 2018 Siberian Transport Forum - TransSiberia, TS 2018; Novosibirsk ExpoCentreNovosibirsk 239

[31] Kots M and Chukanov V 2019 U-Net adaptation for multiple instance learning Journal of Physics: Conference Series 1236

[32] Shariati F, Hosseinlou1 S and Rud’ V 2019 Automatic lung segmentation method in computed tomography scans Journal of Physics: Conference Series 1236

[33] Chukanov V and Bergner F 2016 Combination of voxel-based and projection-based methods in terms of convergence for CT reconstruction Medical Physics 43 2828-2834

[34] Skvortsov A N, Savchenko E A, Velichko E N, Madzhinov A R and Nezhinskikh S S 2019 Journal of Physics: Conference Series 1236 1042

[35] Savchenko E A, Kuznetsova O B, Andryakov A A, Savchenko E Y and Musakulova Z A, 2019 Journal of Physics: Conference Series 1236 1039

[36] Velichko E, Savchenko E, Nepomnyashchaya E, Dyubo D and Tsybin O 2018 Internet of Things, Smart Spaces, and Next Generation Networks and Systems 649