Deep learning for brands object detection and recognition in images

Vladimir Golovko[1], Egor Mikhno[2], Alexander Kroschenko[3], Sergei Bezobrazov[4]
1) Brest State Technical University and Państwowa Szkoła Wyższa im. Papieża Jana Pawła II, Moskovskaja 267 Brest, gva@bstu.by, iit.bstu.by
2) Brest State Technical University, Moskovskaja 267 Brest, dzinushi.kun@gmail.com, iit.bstu.by
3) Brest State Technical University, Moskovskaja 267 Brest, kroschenko@gmail.com, iit.bstu.by
4) Brest State Technical University, Moskovskaja 267 Brest, bescase@gmail.com, iit.bstu.by

Abstract: *In this paper, we investigate applying several well-known models to the task of brands detection in images.*
We implemented comparison of most effective and widely used architectures as Faster R-CNN (ResNet50/101), SSD u YOLO. Received results confirm the effectiveness of applying Faster R-CNN to any sets of images. However, it is necessary to note the resource-intensiveness of this architecture and its unsuitability for solving problems, in which an important criterion of efficiency is the time for performing the analysis. The SSD and YOLO models do not offer advantages in the detection of small and medium-sized objects, but can be successfully used as part of mobile detection systems that are limited in their hardware capabilities. In addition, these neural network architectures perform processing faster than Faster R-CNN and can be considered as basic models for detecting and segmentation of objects in images and video in real time.

Keywords: Deep neural network, image detection, image classification, beer brands.

1. PROBLEM

The task of object detection on the images is one of the most actively studied tasks of artificial intelligence. Objects search and counting their number on images or video is a great important task for a business. Routine tasks for the manual assessment of the number of various types of goods take a considerable part of the work-time of specialists. Therefore, the application of recent research in the field of deep convolutional neural networks for the tasks of detection and classification can help to automate such routine work. In this work, we investigated various models for the detection of goods of certain trademarks on images. The obtained results allow us to talk about the degree of applicability of various models to the proposed detection problem.

2. EXISTING SOLUTIONS

Object detection is one of the most popular research areas in machine learning for recent years. To solve this problem, traditional methods based on the use of SIFT (Scale-Invariant Feature Transform) [1] and SURF (Speeded Up Robust Features) feature tags [2] were actively used. The SIFT method is based on extracting key points from a set of objects of interest and comparing them with new analyzed images. SIFT allows you to detect an object in the presence of noise and partial overlap. The SURF method is based on SIFT, but at the same time, it has a greater work speed [3]. Both methods have high mathematical complexity and, in general, have low generalizing ability in comparison with modern methods based on using convolutional neural networks [4].

The advances made in learning deep neural networks have influenced the methods used in the detection of objects. So, ideas and approaches based on the use of various neural network architectures began to actively develop. In 2014, R-CNN [5] was proposed, in 2015 - Fast R-CNN [6], a feature of which was the use of a special ROI-layer, which made it possible to speed up the network. Following it, the Faster R-CNN architecture was developed [7], which differs from Fast R-CNN in the presence of a special RPN network (Region Proposal Network), whose main task is to highlight areas of applicants. Such changes made it possible not only to speed up the work of the network but also to get better indicators of the generalizing ability compared to Fast R-CNN. In 2016, YOLO [8] and SSD [9] architectures were proposed.

All of ours architectures are divided into two main categories:

1. Methods with a preliminary selection of candidates (R-CNN, Fast R-CNN, Faster R-CNN);
2. One-way methods (one-look), which include SSD, YOLO, YOLO9000.

The peculiarity of the second group of methods is the detection of objects in the image in one pass (one-look), without the need to solve two independent tasks, namely the localization of the object and its classification. For the methods of the first group, these two problems are solved by separate parts of the neural network architecture or even by separate methods (R-CNN).

All models for detecting objects in images are based on the use of a previously trained deep convolutional neural network. Most often these are neural networks for classification without the last fully connected layer. Then the network is trained on new data. Thus, the pre-trained network plays the role of a "supplier" of features for the layers performing detection.

3. DATASET

We used photographs from supermarkets, provided by LeverX [10], as initial data. Examples of images (RGB) used for training are shown in Fig. 1.

We used a general sample of 783 photographs, with 650 images of this sample used for training and the remaining 133 images for testing. In the marking process, ten of the most frequently encountered beer brands were identified. Sample preparation consisted of manually sorting the images with the definition for each of the characteristics of rectangular areas that include goods (height, width, coordinates of the upper left corner). On one image there can be several areas of interest to us (for

the case of several boxes of goods located one on one).

Fig.1 – Example of images from the training set

When preparing a training set for solving a detection the problem, the use of bounding boxes to select some objects does not seem appropriate since objects can have a complex shape far from a rectangular one (Fig. 2). This is explained by the fact that initially, three-dimensional objects (such as boxes) are difficult to place in a rectangular area without including unnecessary elements (such as fragments of other boxes, a background image, etc.). For such objects, it is possible to obtain acceptable detection results if, when marking them out, to focus on the image of the trademark, and not on the container on which it is located.

4. PROPOSED SOLUTION

To solve the problem of detection, we used several different architectures of deep neural networks. All of them showed acceptable results in solving the detection problem. The analysis was carried out for the following architectures: Faster R-CNN, based on the ResNet-50/101 classifier [11], SSD and YOLO.

The Faster R-CNN model consists of three parts (Fig. 3). The first part is the ResNet-50 (ResNet-101) classifier, pre-trained on a COCO sample [12]. The second part is the RPN network that generates the candidate regions. Finally, the third part is the detector, which is represented by additional fully connected layers that generate the coordinates of bounding boxes containing the desired objects, and class labels for each such area. A key feature of the model is the RPN-network, to the input of which the feature maps obtained by the preceding convolutional layer are fed. Due to this, the generation of applicants is faster than using the original full-size image.

5. MODELS EVALUATION

To assess the effectiveness of the trained models, we used the mAP (mean average precision metric). This metric is the most frequently used for assessing the quality of detection models. It is used in conjunction with its modifications calculated for various threshold values of IoU (Intersection over Union, a quantity called the Jaccard measure).

The value of IoU is calculated by the following way:

$$IoU = \frac{S_{ground_{true}} \cap S_{box}}{S_{ground_{true}} \cup S_{box}} \qquad (1)$$

where $S_{ground_{true}}$ – area of reference box that is used to mark the training set, S_{box} – area of the box generated by the model.

Fig.2 – Example image with difficult objects form

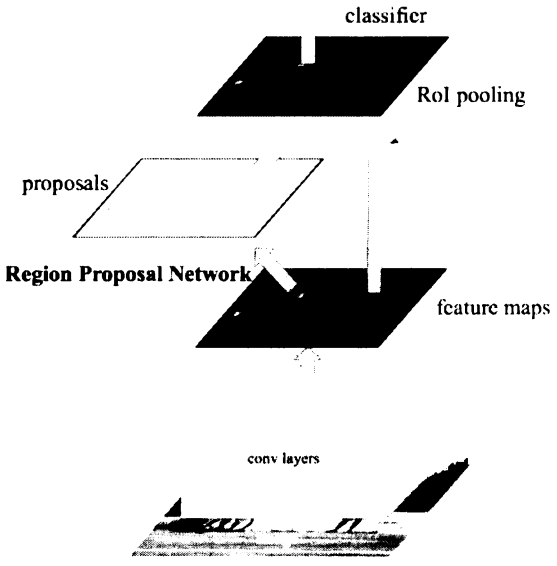

Fig.3 – Structure of Faster R-CNN [7]

As you know, the proportion of correct detections in the total number of detections obtained by the neural network is P and calculated as follows:

$$P = \frac{TP}{TP + FP} \qquad (2)$$

where TP – the number of true-positive, FP – the number of false-positive detection results.

Regarding the task of detecting objects, the number of TP determines the total number of bounding boxes for which the IoU value, calculated relative to the true areas (Ground-true box), is greater than a certain threshold (most often the threshold value is chosen 0.5). Thus, if the IoU value for such predicted area greater than 0.5, then detection is considered as true positive. If there are several detections for this true region, then one detection is selected with the largest IoU value, and the rest are considered as FP (hard non-maximum suppression).

The averaged value for all sensitivity values gives AP:

$$AP = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \qquad (3)$$

where N – the number of sensitivity values calculated at regular intervals.

The mAP value is obtained by averaging the AP over all the object classes considered.

6. RESULTS

All tested models were trained at a different number of iterations. In average, an acceptable result was achieved after 5,000 iterations of training. We used a general approach: the model was trained for 1000 iterations, after which it was tested, then the process was repeated for the next 1000 iterations. If after next stage of learning the results didn't improved or become worse, the process was completed.

Table 1 shows the results obtained for various architectures and optimization methods. The table below shows that the best results were obtained by Faster-RCNN architecture (ResNet-50/101). The rest of the architectures (such as YOLO and SSD) showed the worst detection results, regardless of the training method used and the number of learning iterations. This is explained by the fact that SSD and YOLO as a whole have a poor ability to detect small objects in an image [13, 14] - feature maps for such architectures have low resolution (usually 38x38 or 19x19). For the considered detection problem, this is of critical importance, since all the images from the sample have a sufficiently high resolution, but the relative size of the objects is small. However, Faster-RCNN architecture is more resource intensive than SSD and YOLO.

Table 1. Detection results for different architectures

Architecture	Number of training iteration	Batch size	Optimizer	mAP
Faster R-CNN (ResNet50)	7150	1	Adam	0.841
Faster-RCNN (ResNet101)	8450	1	Adam	0,824
SSD	30000	4	RMSProp	0,675
YOLO	2000	8	Adam	0,628

Figures 4 show the results of the detection of single and several products respectively.

Fig.4 – Detection results for individual objects classes

7. CONCLUSION

This article discusses the use of various models to solve the problem of detecting goods of different brands in the image.

A comparative analysis of the most efficient and widely used neural network architectures Faster R-CNN (ResNet50 / 101), SSD and YOLO have been carried out. The results confirm the effectiveness of applying the Faster R-CNN architecture to any image samples. However, it is necessary to note the resource intensity of such architectures and their unsuitability for solving problems, in which the analysis time is an important criterion of efficiency. At the same time, SSD and YOLO models can be successfully used as part of mobile detection systems, limited in their hardware capabilities. In addition, these neural network architectures perform faster Faster R-CNN processing and can be considered as basic models for detecting and segmentation of photo and video images in real time.

8. REFERENCES

[1] Lowe, D. Object recognition from local scale-invariant features / D. Lowe // Proceedings of the International Conference on Computer Vision. – 1999. – Vol. 2. - P. 1150–1157.

[2] Bay, H. SURF: Speeded Up Robust Features / H. Bay, T. Tuytelaars, L. Van Gool // Proceedings of the ninth European Conference on Computer Vision. – 2006.

[3] Panchal, P. M. A comparison of SIFT and SURF / Panchal, P. M., S. R. Panchal, and S. K. Shah // International Journal of Innovative Research in Computer and Communication Engineering, 1(2). – 2013. – P. 323–327.

[4] LobnaRagab, S. Object Detection using Histogram and SIFTAlgorithmVs Convolutional Neural Networks / S. LobnaRagab // Academia [Веб-ресурс]. – 2014. – Access mode: http://www.academia.edu/24497785/Object_Detection_using_Histogram_and_SIFT_Algorithm_Vs_Convolutional_Neural_Networks. – Access date: 20.12.2018.

[5] Girshick, R. Rich feature hierarchies for accurate object detection and semantic segmentation / R. Girshick, J. Donahue, T. Darrell, J. Malik // arXiv [Web-resource]. – 2014. – Access mode: https://arxiv.org/pdf/1311.2524v5.pdf. – Access date: 20.12.2018.

[6] Girshick, R. Fast R-CNN / R. Girshick // arXiv [Web-resource]. – 2015. – Access mode: https://arxiv.org/pdf/1504.08083.pdf. – Access date: 20.12.2018.

[7] Ren, S. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / S. Ren, K. He, R. Girshick, J. Sun // arXiv [Web-resource]. – 2016. – Access mode: https://arxiv.org/pdf/1506.01497.pdf. – Access date: 20.12.2018.

[8] Redmon, J. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // arXiv [Web-resource]. – 2016. – Access mode: https://arxiv.org/pdf/1506.02640.pdf. – Access date: 20.12.2018.

[9] Liu, W. SSD: Single Shot MultiBox Detector / W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. C. Berg // arXiv [Web-resource]. – 2016. –

Access mode: https://arxiv.org/pdf/1512.02325.pdf. – Access date: 20.12.2018.

[10] Hire SAP integrator, long-term SAP service provider | LeverX // LeverX [Web-resource]. – 2018. – Access mode: https://leverx.com. – Access date: 20.12.2018.

[11] Kaiming, He Deep Residual Learning for Image Recognition / H. Kaiming, Z. Xiangyu, R. Shaoqing, S. Jian // arXiv [Web-resource]. – 2015. – Access mode: https://arxiv.org/pdf/1512.03385.pdf. – Access date: 20.12.2018.

[12] Lin, T. Microsoft COCO: Common Objects in Context / T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár // arXiv [Web-resource]. – 2015. – Access mode: https://arxiv.org/pdf/1405.0312.pdf. – Access date: 20.12.2018.

[13] Huang, J. Speed/accuracy trade-offs for modern convolutional object detectors / J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy // Computer Vision and Pattern Recognition. – 2017. – P. 7310–7319.

[14] Hui, J. What do we learn from single shot object detectors (SSD, YOLOv3), FPN \& Focal loss (RetinaNet)? / J. Hui // Medium.com [Web-resource]. – 2018. – Access mode: https://medium.com/@jonathan_hui/what-do-we-learn-from-single-shot-object-detectors-ssd-yolo-fpn-focal-loss-3888677c5f4d. – Access date: 20.12.2018.