Deep Convolutional Neural Network for Recognizing the Images of Text Documents

Vladimir Golovko[1,2][0000-0003-2615-289X], Aliaksandr Kroshchanka[1][0000-0003-3285-3545], Egor Mikhno[1][0000-0002-7667-7486], Myroslav Komar[3][0000-0001-6541-0359], Anatoliy Sachenko[3,4][0000-0002-0907-3682], Sergei Bezobrazov[1][0000-0001-6436-2922], Inna Shylinska[3][0000-0002-0700-793X]

[1] Brest State Technical University, Brest, Belarus
[2]Państwowa Szkoła Wyższa im. Papieża Jana Pawła II, Biala Podlaska, Poland
[3]Ternopil National Economic University, Ternopil, Ukraine
[4]Kazimierz Pulaski University of Technology and Humanitie, Radom, Poland
vladimir.golovko@gmail.com[1,2], mko@tneu.edu.ua[3], sachenkoa@yahoo.com[3,4]

Abstract. A comparative analysis of various methods and architectures used to solve the problem of object detection is carried out. This allows so-called one-way neural networks architectures to provide high quality solutions to the problem. A neural network algorithm for labeling images in text documents is developed on the basis of image preprocessing that simplifies the localization of individual parts of a document and the subsequent recognition of localized blocks using a deep convolutional neural network. The resulting algorithm provides a high quality of localization and an acceptable level of subsequent classification.

Keywords: Object Detection, Deep Convolutional Neural Network, Labeling Images, Image Preprocessing, Text Image.

1 Introduction

Deep learning is a very effective technique in the domain of machine learning and has been successfully applied to many problems referring to artificial intelligence, namely object detection, natural language processing, data visualization, etc. Different techniques for deep neural networks learning exist [1-4]. We address to the object detection in text images using deep learning techniques in this paper.

To detect objects in images it is necessary to select separate blocks of an image referring to certain predefined classes. A model that performs such an operation receives an image at the input, and gives back the coordinates and dimensions of rectangular areas including the objects being searched for as well as the probability of referring the included object to a certain class.

The solution to such a problem is one of the challenges in computer science. In fact, due to this functionality, it is possible to analyze photos and video images in real time by placing labels on certain objects and performing predefined processing

operations. At the same time, it is necessary to distinguish between the object detection problem and the semantic segmentation problem that is actually the classification of each image pixel.

The object detection task can be logically divided into two subtasks – the localization of the object and its classification. Many existing approaches to detecting objects in images allow us to combine these two distinct stages in a single neural network, which performs both tasks simultaneously and gives the final result at the output (Fig. 1). This makes it possible to solve the problem much faster and receive a result with the information about all the detected objects without the need for their sequential processing. However, one should not completely abandon the classical approaches as there are tasks solving which by such methods provide better results.

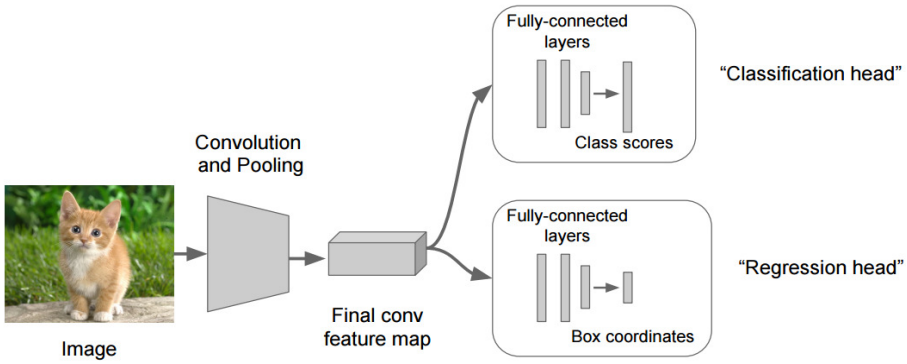


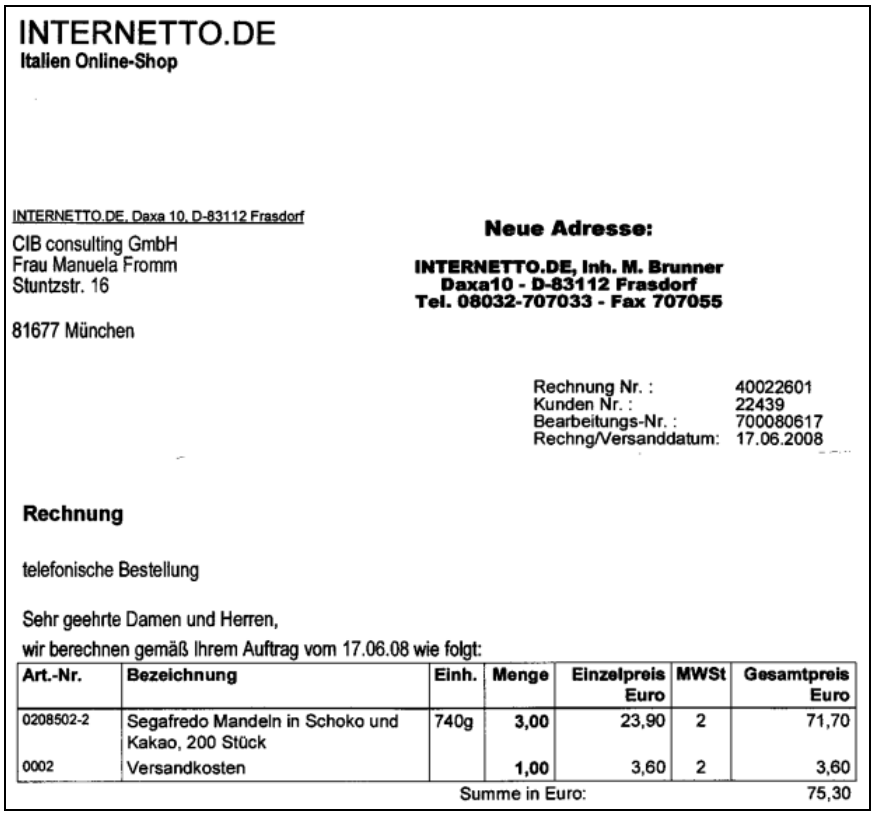
Fig. 1. General view of the neural network model used to solve the problem of detecting objects in an image [5]

The assessment of the quality of the solution to the object localization problem is performed by calculating the IoU (Intersection over Union) metric [6].

We had to solve the problem of detecting objects in documents represented by images. Actually, for such a case, the problem of detecting objects is reduced to the task of labeling electronic document, highlighting its components. An example of the analyzed document from the Doxima7000 sample provided by CIB Software [7] is shown in Fig. 2.

It can be seen in the presented image that the document consists of certain logical units (in particular, company logo, table, text, bank data, address, etc.) that can be detected and further processed (for example, text blocks can be recognized converting them into a format that can be more easily analyzed and interpreted using a computer.

The Doxima7000 data set consists of approximately 7000 documents in German. These documents can be classified into 3 main categories: invoices, receipts and business cards. The distribution of individual categories of documents varies greatly. So, the most numerous is the group of "Text" objects, and the less common are objects of the "Sign" and "Contacts" classes. Comparative distribution of objects of different classes in the first 100 documents of the data set can be seen in Fig. 3.



INTERNETTO.DE
Italien Online-Shop

INTERNETTO.DE, Daxa 10, D-83112 Frasdorf
CIB consulting GmbH
Frau Manuela Fromm
Stuntzstr. 16

81677 München

Neue Adresse:

INTERNETTO.DE, Inh. M. Brunner
Daxa10 - D-83112 Frasdorf
Tel. 08032-707033 - Fax 707055

Rechnung Nr. : 40022601
Kunden Nr. : 22439
Bearbeitungs-Nr. : 700080617
Rechng/Versanddatum: 17.06.2008

Rechnung

telefonische Bestellung

Sehr geehrte Damen und Herren,
wir berechnen gemäß Ihrem Auftrag vom 17.06.08 wie folgt:

Art.-Nr.	Bezeichnung	Einh.	Menge	Einzelpreis Euro	MWSt	Gesamtpreis Euro
0208502-2	Segafredo Mandeln in Schoko und Kakao, 200 Stück	740g	3,00	23,90	2	71,70
0002	Versandkosten		1,00	3,60	2	3,60
				Summe in Euro:		75,30

Fig. 2. Fragment of a document from the Doxima7000 data set

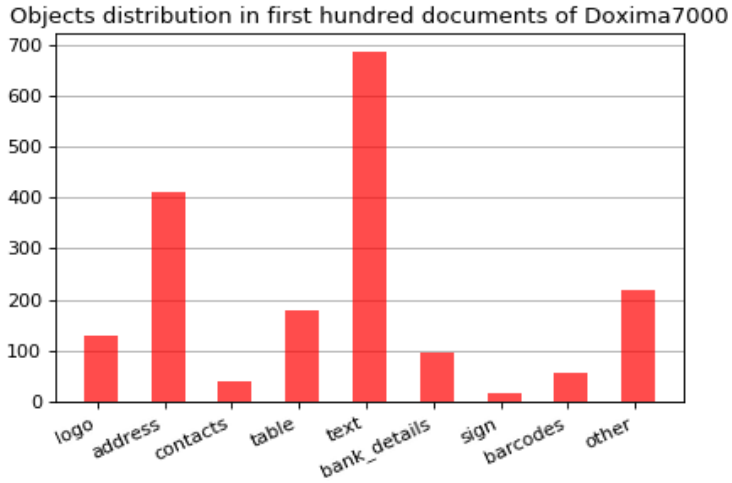


Fig. 3. Objects distribution in Doxima7000 data set

This data set was not originally intended for training and testing neural network detection models, therefore it does not include the document labeling. Thus, the labeling was done manually, by highlighting the corresponding block of the document and assigning it to a certain predefined class.

For training and testing classic neural network detection models, data set of 145 and 36 documents, respectively, were used. However, in the proposed neural network model, the labeling and submission of the whole documents was not compliant with the model architecture. Therefore, we used a training data set of 2500 images of individual blocks of documents for neural network learning. This allows us to ensure a balanced sample with respect to the main classes of blocks.

2 Application of Conventional Architectures

2.1 Faster R-CNN

The first neural network model used by us for the detection of individual blocks of documents was the Faster R-CNN [8]. This model consists of three parts (Fig. 4). The first part is the ResNet-50 (ResNet-101) classifier, which was trained on the COCO data set [9]. 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 rectangular areas containing the desired objects and class labels for each of such areas. The speed and efficiency of the analysis is significantly influenced by the RPN network, at the input of which the feature maps obtained by the preceding convolutional layer are fed. Due to this, candidate regions generation is faster than the use of the original full-size image.

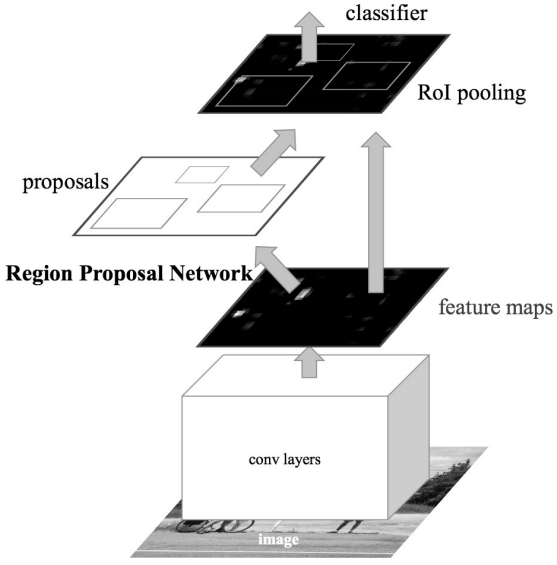


Fig. 4. Faster R-CNN structure [8]

2.2 SSD (Single-shot detector)

The SSD model [10], as well as YOLO [11], belongs to the category of single-pass methods that allow solving the problem of detecting objects within a single network.

Schematic representation of the architecture is shown in Fig. 5.

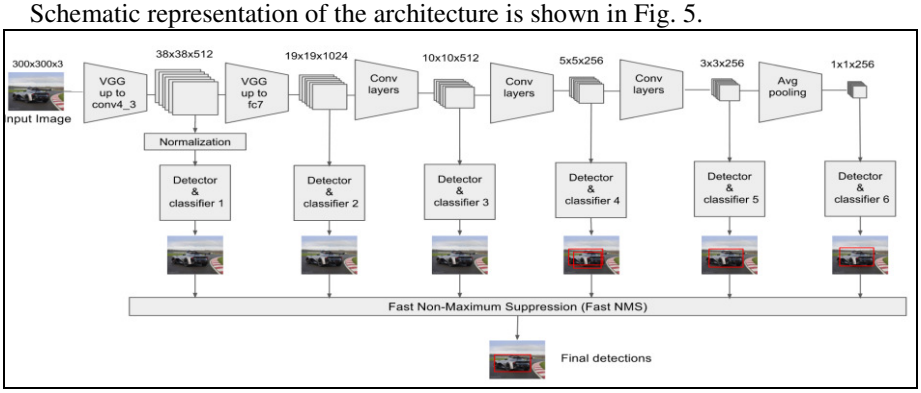


Fig. 5. SSD architecture diagram

The main features of this model:

- It accumulate information about objects position and scale through subsequent convolutional layers. Each of this layers detects objects of specific size. (Fig.6).
- The pre-trained network (VGG or ResNet) is used as the base element, which is converted to a fully CNN (FCN).
- Non-maxima suppression is used to decrease the number of detected boxes during network operation.
- Each feature map cell creates a group of default boxes (or anchors), differing in scale and aspect ratio.
- The model is trained in order for each anchor to correctly predict its class and offset.

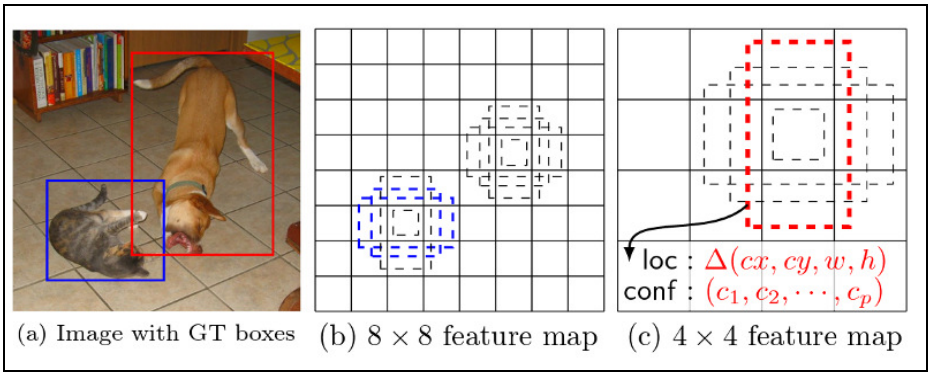


Fig. 6. Localization of objects in feature maps of different sizes [8]

The results of applying the SSD architecture to solving the task are shown in Fig. 7.

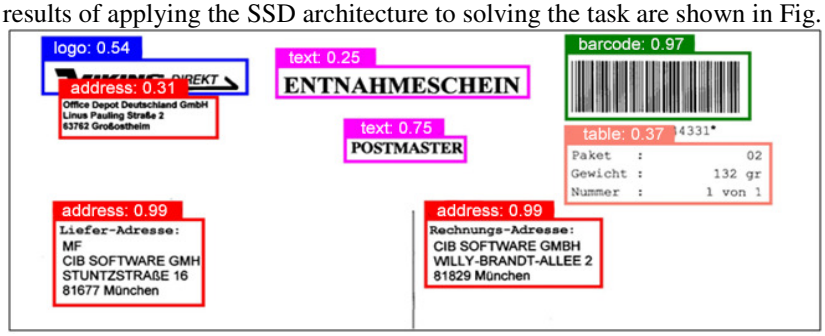


Fig. 7. Example of document labeling, made by SSD model

3 The Proposed Neural Network Approach to Labeling the Text Images

In addition to reviewing and studying standard solutions to the object detection problem, we proposed an original approach based on the R-CNN method that includes two processing steps [12, 13]. At the first stage, the selection of the regions of interest is carried out; it is preferable when working with the text data. At the second stage, the regions were classified using the classical convolutional network. Let us dwell on the description of each processing stage.

I. The following operations are applied to the original image:

- Median filter – to remove noise in the original document, associated with non-ideal conditions of scanning a document, printing, etc.
- Box filter – a linear filter used to create a blur effect (necessary for suppressing small details and highlighting regions with the same type of content).
- Applying the threshold function for the formation of continuous regions.
- Selecting the contour areas and localization of text blocks.

The first 3 of these operations are shown in Fig. 8.

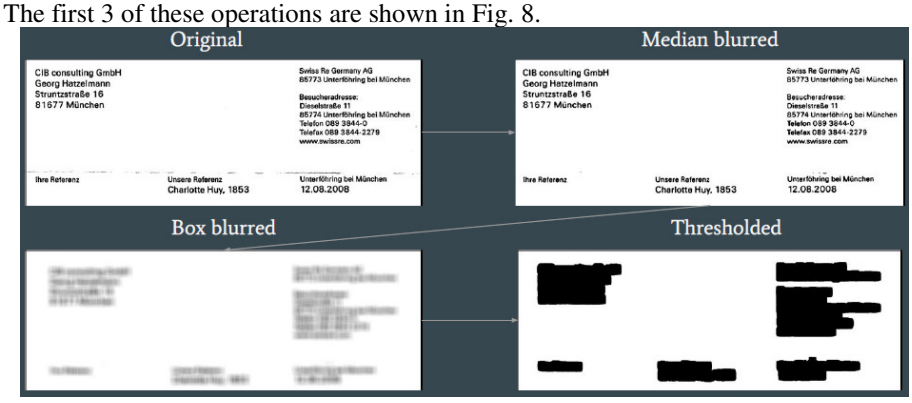


Fig. 8. Stages of pre-processing of the original image for localization of the text blocks

After performing the above-mentioned operations, we get a set of rectangular areas containing text blocks (Fig. 9).

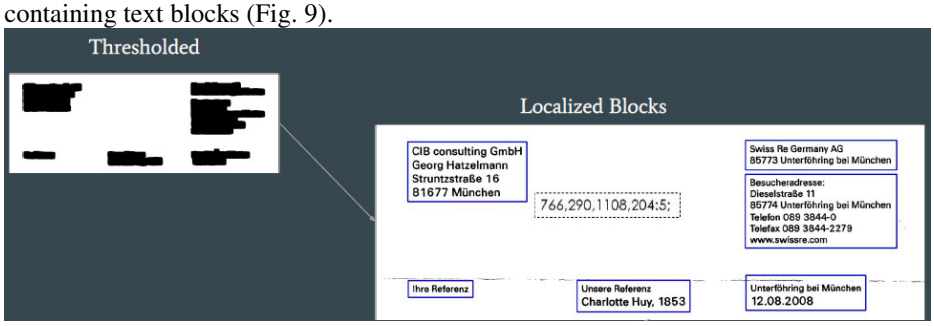


Fig. 9. Results of object localization

II. Training convolutional neural network. For recognition of blocks obtained at the first stage, a convolutional neural network is trained.

At this stage, we have created a training sample, consisting of about 2500 images, divided into 7 classes (logos, bar- and QR-codes and signatures, etc.). The neural network selected as a working model is shown in Fig. 10.

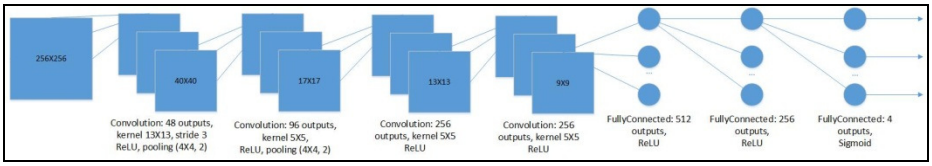


Fig. 10. Convolutional neural network for the classification of the text blocks

After completing the training, the obtained test results are presented in Table 1.

Table 1. Test results (correctly recognized objects in percentage %)

Logos, %	Bar codes, %	Signatures, %	Other, %
97.65	100	97.76	99.29

Thus, rather high efficiency indicators for the trained classifier were obtained. Results of identifying certain types of documents are shown in Fig. 11.

Obtained results can be used by neural network immune detectors for identification and classification of computer attacks and malicious application detection in Android OS [14].

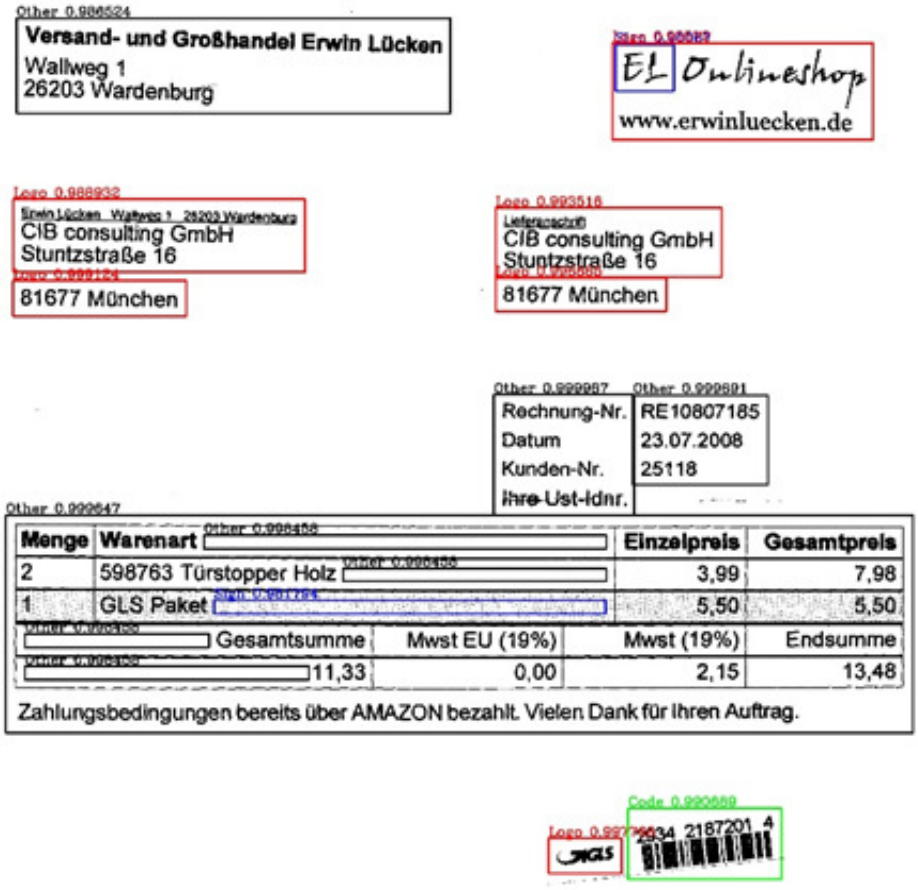


Fig. 11. Results of object detection system

4 Analysis of Results

In order to analyze and compare the different neural network approaches we have used the standard metrix *mAP*.

The *mAP* (mean average precision metric) [15] in context of this work was used to evaluate the quality of detection of document parts. Sometimes *mAP* is used with its modifications computed for various values of *IoU* (Intersection over Union, Jaccard index). *IoU* is calculated in the following way (Fig. 12):

$$IoU = \frac{S_{ground\_true} \cap S_{box}}{S_{ground\_true} \cup S_{box}} \tag{1}$$

where S_{ground_true} determines the area of the reference block, which is used for labeling the testing set, and S_{box} is the area of the detected block.

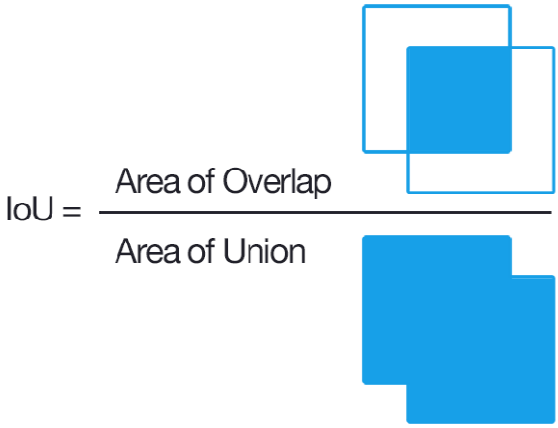


Fig. 12. Calculating *IoU* metrics

For objects detection task, the number true-positive detection results defines based on the total number of rectangular blocks for which the value of *IoU* exceeds some threshold (usually the threshold of 0.5 is chosen). *TP*-results are calculated for real block (ground-true block). If real block has several detections, for it is selected only one block with largest value of *IoU* and other blocks are considered as *FP*.

The averaged value for all values of recall gives the *AP*:

$$AP = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{2}$$

where *N* is the number of equally spaced recall values.

The value of *mAP* is obtained from *AP* by averaging over all classes of objects.

The results of object detection in text images for different approaches are shown in the Table 2.

Table 2. Comparison of different approaches

Model	mAP	FPS
Faster R-CNN	0,9184	10
SSD (Inception v2)	0,8321	23
Preprocessing + CNN	0,8955	21

As can been seen from the Table 2 the proposed approach has the best results concerning Frames per Second (FPS) and acceptable accuracy of object detection.

5 Conclusion

The neural network algorithm for labeling images in text documents based on image preprocessing was developed. The algorithm simplifies the localization of individual parts of a document and the subsequent recognition of localized blocks using a deep convolutional neural network. The resulting algorithm provides a high quality of localization and an acceptable level of subsequent classification. In addition, a comparative analysis of various methods and architectures used to solve the object detection problem was carried out. This allows so-called one-way neural networks architectures to provide high quality solutions to the problem.

References

1. LeCun, Y., Bengio, Y., Hinton, G. Deep learning, Nature, 521 (7553), 436–444. (2015).
2. Hinton, G., Salakhutdinov, R. Reducing the dimensionality of data with neural networks, Science, 313 (5786), 504–507. (2006).
3. Bengio, Y. Learning deep architectures for AI, Foundations and Trends in Machine Learning, 2(1), 1–127. (2009).
4. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y. Object recognition with gradientbased learning, Shape, Contour and Grouping in Computer Vision, 1681, 319-345. (1999).
5. Object Localization and detection, https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/object_localization_and_detection.html, last accessed 2019/03/07.
6. Intersection over Union (IoU) for object detection, https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection, last accessed 2019/03/07.
7. Cib software, https://cib.by, last accessed 2019/03/07.
8. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, https://arxiv.org/pdf/1506.01497.pdf, last accessed 2019/03/07.
9. Lin, T. Maire, M. Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollár, P. Microsoft COCO: Common Objects in Context, https://arxiv.org/pdf/1405.0312.pdf, last accessed 2019/03/07.
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A. C. SSD: Single Shot MultiBox Detector, https://arxiv.org/pdf/1512.02325.pdf, last accessed 2019/03/07.
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.,You Only Look Once: Unified, Real-Time Object Detection,https://arxiv.org/pdf/1506.02640.pdf, last accessed 2019/03/07.
12. Kroshchenko, A., Golovko,V., Bezobrazov, S., Mikhno, E., Khatskevich, M., Mikhnyaev, A., Brich, A. Deep trainingfor detecting of objects at images of documents, Vesnyk Brest State Technical University, 5(107), 2–9. (2017) (In Russian).
13. Golovko, V., Mikhno, E., Brich, A., Sachenko, A. A Shallow Convolutional Neural Network for Accurate Handwritten Digits Classification, Communications in Computer and Information Science, 673, 77–85. (2017).
14. Komar, M., Sachenko, A., Bezobrazov, S., Golovko, V. Intelligent Cyber Defense System Using Artificial Neural Network and Immune System Techniques, Communications in Computer and Information Science, 783, 36-55. (2017).
15. Oksuz, K., Cam, B. C., Akbas, E., Kalkan, S. Localization Recall Precision (LRP): A New Performance Metric for Object Detection, https://arxiv.org/pdf/1807.01696.pdf, last accessed 2019/03/07.