Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Research of YOLO Architecture Models in Book Detection

Maria Kalinina
Department 316
Moscow Aviation Institute (National Research University)
Moscow, Russia
Ptaha-96@yandex.ru

Pavel Nikolaev*
Department 316
Moscow Aviation Institute (National Research University)
Moscow, Russia
npavel89@gmail.com

***Abstract*— Deep neural networks are widely used in different fields of human activity, including spheres which are connected with large amount of operations such as data obtaining and processing information from the outside world. This article deals with the creation of the deep convolutional neural network based on the YOLO architecture for book detection in real time. The architecture chosen as the basis of the neural network possesses a number of advantages which make it highly competitive with other models, so it can be considered as the most suitable option for the creation of deep neural network for object detection. Creation of the original dataset and the deep neural network training are described. Several variants of neural networks based on the YOLO architecture are discussed and the results of their comparison are shown. The results obtained during the training of a deep neural network allow us to use it as a basis for further development of the application.**

Keywords—image recognition, object detection, computer vision, machine learning, artificial neural networks, deep learning, convolutional neural networks

I. INTRODUCTION

At the present time neural networks are widely used in various spheres of human society and often act as an assistant in solution of many important problems which is based on the object detection in many cases. Among various types of networks convolutional neural networks (CNN) show the best results in image recognition.

The use of neural networks can greatly facilitate activities in various areas, for example, those that deal with large amounts of data, for instance, while working with a large number of books, which is typical for bookstores, libraries or warehouses. Search of the definite books based on the use of a neural network capable to localize book spines can be more efficient and faster than in case if it is performed manually. On the basis of such a neural network, it is possible to create a specialized application that can act, for example, as an inventory search engine or an independent mobile application.

This article discusses the development of a neural network for detecting books on bookshelves with the use of book spines. Book detection is supposed to be performed in real time. In this regard, it is necessary to build and train a neural network that can recognize a sufficient number of frames per second and at the same time give fairly accurate results.

The neural network should analyze the input image for the presence of books on it and provide the user with information about the detected objects. In other words, at the output the location of the books should be marked on the submitted image with the use of a bounding box ("border" around the spine of the detected book) located on the book spine.

The main tasks of this work are to create and compare several models of the CNN for detecting books and their further training and comparison of the results obtained in order to identify the optimal model.

II. YOLO CONVOLUTIONAL NETWORK

Nowadays deep convolutional networks successfully act as the systems of deep learning and show good results in different tasks, for instance, in case of image classification, object detection and segmentation.

The typical architecture of CNN is based on the mixture of convolutional and pooling layers. While passing through it picture transforms into the feature map which consequently goes to fully-connected layers.

You Only Look Once (YOLO) architecture refers to the type of one-shot detectors [1-3]. This type of networks is notable for remarkable speed of work, however precision of obtained results due to the single passing of the image through the network can be a little lower. It can be quite useful to put in practice this type of network as a basement for software if it deals with the real time functioning and allows to recognize sufficient quantity of frames.

According to the results of researches [3] the last modification of YOLO – YOLOv3 is turning out to demonstrate the highest speed of work and precision in comparison with other networks of the same type of detectors. YOLO can be used for solving problems of different types, for example, for ship detection [4], vehicle traffic analysis [5], traffic signposts [6], bridge damage detection [7], in case of medical tasks [8], etc.

One of the important features of YOLO-based networks is the use of grid of a definite size which superimposes the input image and divides it into a number of cells. Each cell is given an array of predicted values: x and y – the coordinates of the bounding box (upper left and lower right points) or the coordinates of the center of the bounding box, w and h – width and height of the bounding box and also C – the probability of the class to which object detected in the image belong. The input image of the neural network is divided into S cells. For

each object in the picture there is a cell which is responsible for its prediction (the cell where the center of the object falls in). For each cell prediction is performed for B bounding boxes and C probabilistic estimates of classes. Thus, the terminal output takes the shape of S x S x (B x 5 + C)-dimensional tensor.

After receiving the input image the network performs necessary operations to split it into cells. During this process the non-max suppression algorithm is applied with the purpose of elimination of unnecessary predictions.

The input data for the algorithm include a list with the following elements: values of the specified bounding boxes, the probability of the presence of the desired object on the image and the overlap thresholds. As a result the output data present a list of predictions which have successfully passed some kind of filtering which can be described as a set of definite actions: the preliminary created empty list is filled with the prediction with the maximum value of the probability of the object's presence in the image, then Intersection Over Union (IOU) values for this and all other predictions are calculated. If IOU values exceed the value of specified overlap threshold the predicted bounding box it belongs to is deleted. In the end, the frame coordinates (or center coordinates with width and height) are obtained in the last convolutional layer.

Our network is intended to be used for detection of objects of a single class. It means that the output data of our network will be presented in the form of a massive which contains 4 values.

According to [1], the YOLO network has a number of advantages such as very high speed of work, which allows it to recognize up to 35 frames per second (according to the speed of the latest version of YOLO on the COCO data set was 35 fps [9]), possibility to operate with the whole image at once, not with its individual parts and to study generalized representations of objects. All these peculiarities cause better work of the network on new data and fewer mistakes while separating the desired objects from the background.

Batch-normalization, a method mostly oriented on normalizing input layers by scaling the dataset, can be applied in order to increase the stabilization and performance of neural networks. Batch-normalization is based on preliminary processing of data that the network will later operate on, to a state where the set of values submitted to the input of the neural network which is characterized by a zero mathematical distribution and the variance is 0.

Leaky ReLU which a modified version of the normal ReLU function is often used as an activation function in YOLO networks. This modification allows successfully use the activation function in cases where conventional ReLU fails, for example, if the neural network training speed is too high.

III. Network for book detection

YOLO-based network for book detection was trained from scratch, without using ready-made scales. A specially created dataset based on 500 images was used in process of network training.

Each image corresponds to one of the markup files in xml format with the total information for the objects of the desired class: class labels, coordinates of the bounding box for each object, name of the file, height and width of the image. The coordinates of the bounding box are xmin and ymin for the upper-left corners and xmax and ymax for lower-right ones. Image markup was performed in a form similar to the markup form of the PascalVOC dataset.

In our case the used data set is not large enough and it is sufficiently homogeneous so we decide to apply the augmentation technique to expand the initial data volume contained in the dataset we have already created. This means that new elements were artificially generated on the base of the original dataset. Augmentation was performed on randomly selected images from the dataset and included the actions such as scaling, rotation, flip, cropping distortion of colours and so on.

To assess the quality of neural networks various metrics are used. Among them there is one of the most popular metrics known as average precision (AP). It is used in various tasks, for example, in cases when we need to evaluate the quality of ranking, classification, or objects detection. The AP for a series of queries can be defined as the average value of the average accuracy estimates for each query [10]. In our case, the metric determines the percentage of correctly detected objects. Mathematically, AP is a calculation of the values of p(r) – the recall function-in the interval of values r = (0; 1).

For the programming language basement for our networks we use Python 3.6. Creation of CNN structure for book detection was done with the use of Python libraries for deep learning – Keras 2.2.4 and TensorFlow 1.12.0.

The dataset on which the deep neural network was trained consists of 500 images with 5,245 book spines depicted on them. Further, the dataset was divided in certain percentage: into training (60% of pictures), validation (10%) and test (30%) sets. Thus, 300 pictures with frames were included in the test set, 50 in the validation set, and 150 in the test set. Moreover, the number of frames in the markup for each picture was individual.

In the process of creating a neural network, several variants of the network structure were tested, including one close to Tiny YOLOv3. Tiny YOLOv3 is a modification of YOLO architecture which is characterized by 2 outputs in comparison with standard 3 outputs in full YOLO networks version. It also has a simplified layer structure which makes Tiny YOLOv3 networks more adaptable for functioning on a mobile platform.

For YOLOv3 architecture networks there are 3 outputs, each is responsible for recognizing objects of a certain size; so one of the outputs with the highest resolution is responsible for detecting small objects, the output with the lowest resolution detects objects that are larger. Accordingly, an output with an average resolution value recognizes medium-sized objects.

IV. Results

In addition to the standard YOLOv3 with 3 outputs, we also tested the modified YOLOv3 network with 2 outputs, which are typical for Tiny YOLOv3, however, in this case

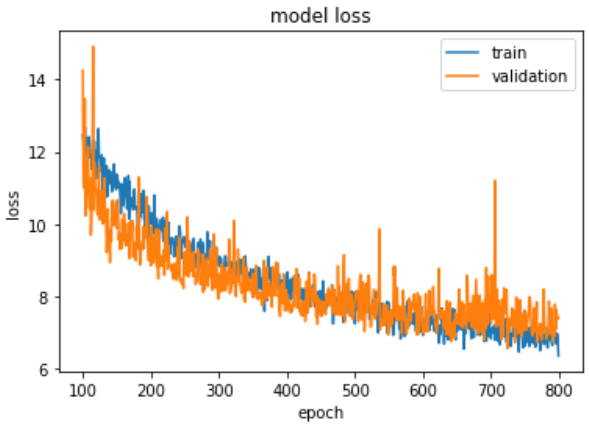

Fig. 1. Changes in error values in training and validation datasets for YOLOv3 Model

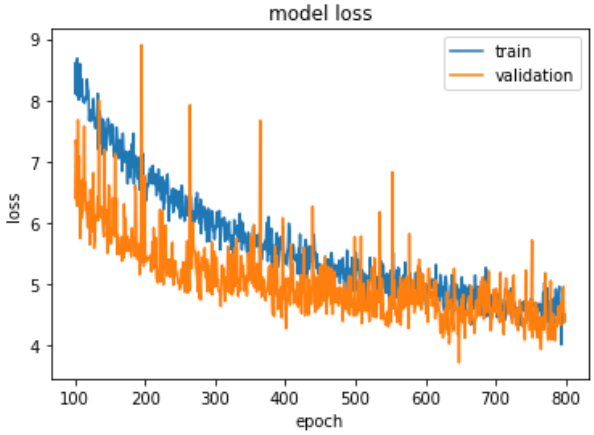

Fig. 2. Changes in error values in training and validation datasets for YOLOv3 Model with 2 outputs

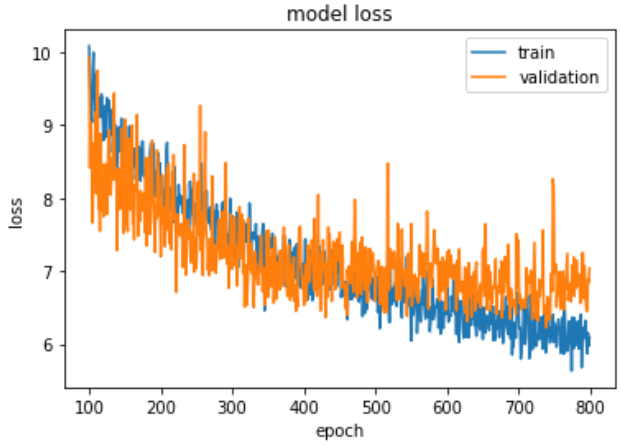

Fig. 3. Changes in error values in training and validation datasets for Tiny-YOLOv3 Model

there is high probability of overfitting, while the standard network can be trained further in order to improve the results.

All of the deep neural networks models were trained during 800 epochs. We trained our network models for book detection on the GPU Nvidia GeForce GTX 1060 6GB with following set of network parameters:

Fig. 4. Example of the neural network work in book detection for YOLOv3 Model

Fig. 5. Example of the neural network work in book detection for YOLOv3 Model with 2 outputs

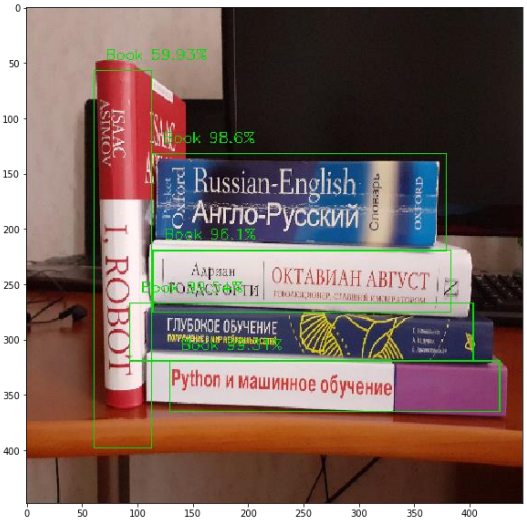

Fig. 6. Example of the neural network work in book detection for Tiny-YOLOv3 Model

1. size of input image – 448x448x3;
2. optimizer – Adam;
3. the learning rate – 0,0001;
4. batch-size – 2 samples;
5. metric for assessing the correctness of the class definition – AP.

During training of the deep neural networks the error function was minimized. The error function can be defined as the sum of errors from each output, so in the models with two outputs the error is less than in the models with three once. Thus, it is necessary to evaluate the network efficiency by the AP value.

The training graphs with the obtained error values for 3 different network models – standard YOLOv3 with 3 outputs, YOLOv3 with 2 outputs and Tiny-YOLOv3 are shown in Fig. 1, 2, 3.

Table 1 shows the best values of AP and loss values for 3 datasets – training, validation and testing obtained during network training for each of 3 tested network models.

TABLE I. THE RESULTS OF YOLO-BASED NETWORKS COMPARISON

YOLO model	**Training set**		**Validation set**		**Testing set**	
	AP	***Loss***	***AP***	***Loss***	***AP***	***Loss***
YOLOv3 (3 outputs)	93.74%	7.38	91.76%	6.57	94.43%	6.74
YOLOv3 (2 outputs)	89.29%	4.53	91.07%	3.72	92.35%	4.42
Tiny YOLOv3	84.98%	5.96	88.42%	6.23	88.75%	6.98

Fig. 4, 5 and 6 show the examples of the functioning of the neural networks after training. Above each frame (bounding box with required object) there are the probability values which corresponds to the probability of found object to belong to a given class. We see that the neural networks quite confidently learned to recognize and detect books on book roots. However, the best results were achieved when using the original YOLOv3 architecture structure.

In our case, experiments with reducing of the number of layers did not bring much success, while increasing their number on the contrary led to improved results. However, if we increase the number of layers, the network may become heavier and the speed of its operation may decrease. We plan to discuss this issue in the next papers.

V. CONCLUSIONS

This article is dedicated to the creation of a neural network for book detection on book spines. As a basis for it several models of YOLOv3 architecture which is distinguished by a significant speed of obtaining output data and a relatively small error value were used. In the process of training neural networks a comparison of the quality of their work was made and the best network was selected. The results obtained after training a deep neural network showed that the network has learned how to recognize and detect objects of a given class confidently. In the future we are planning to continue our work on its improvement.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. “You only look once: Unified, real-time object detection,” Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016.

[2] J. Redmon, A. Farhady. “Yolo9000: Better, faster, stronger,” Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517–6525, 2017.

[3] J. Redmon, A. Farhady. “YOLOv3: An Incremental Improvement,” Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[4] Y. Chang , A. Anagaw, L. Chang, Y. Chun Wang, C. Hsiao, W. Lee. “Ship Detection Based on YOLOv2 for SAR Imagery,” Remote Sensing – Open Access Journal, 2 April 2019.

[5] C. Rajesh Babu, G. Anirudh. “Vehicle Traffic Analysis Using Yolo,” Eurasian Journal of Analytical Chemistry, vol. 13, pp. 345–350, 2018.

[6] A. V. Devyatkin, D. M. Filatov. “Neural network traffic signs detection system development,” Proc. International conference on soft computing and measurement, vol. 1, pp. 189-192, 2019 [Nejrosetevaya sistema obnaruzheniya znakov dorozhnogo dvizheniya].

[7] C. Zhang, C.C. Chang, M. Jamshidi. “Bridge Damage Detection using a Single-Stage Detector and Field Inspection Images,” Computer Society, 8 April 2018.

[8] K.S. Kurochka, T.V. Luchsheva, K.A. Panarin. “Localization of human percentages on X-ray images with use of Darknet YOLO,” Doklady BGUIR, Vol. 113, No. 3, pp. 32–38, 2018 [Lokalizaciya pozvonkov cheloveka na rentgenovskih izobrazheniyah s ispol'zovaniem DARKNET YOLO].

[9] “YOLO: Real-Time Object Detection”. Source: https://pjreddie.com/darknet/yolo/

[10] T.C. Arlen. “Understanding the mAP Evaluation Metric for Object Detection,” Medium, 1 March 2018.