Analysis of Work of YOLO v.3 AND YOLO v.2 Neural Networks

Dmitriy P. Ayrapetov[1], Boris Y. Buyanov[2]
Department of Intelligent Control Systems and Automation
Moscow Technical University of Communication and Informatics
Moscow, Russia
[1]dmitry_a95@live.ru, [2]b.buyanov@gmail.com

***Abstract*. Currently, in the framework of industrial use for solving problems related to the safety of production, widely used computer vision systems, which, as a rule, use neural networks. The article represents an example of settings and operation within the software and hardware complexes of DarkNet Yolo V3 and V2 neural networks, providing the ability to quickly recognize objects. This is critical for making decisions automatically.**

***Key words:* neural networks, object recognition, machine learning, mathematical calculations, industrial Internet of Things.**

I. INTRODUCTION

Currently, in cooperation with the work on the development of artificial intelligence, there is an increased demand for the development of artificial systems based on work algorithms similar to neural networks.

Neural networks are models of biological neural networks of the brain, in which neurons are simulated with relatively simple, often identical, elements (artificial neurons).

Among the areas of neural networks application are automation of image recognition processes, forecasting, adaptive control, expert systems creation, associative memory organization, processing of analog and digital signals, synthesis and identification of electronic circuits and systems [1–3].

The development of modern neural networks is oriented to the processing of natural language, i.e. computer analysis of natural language and its synthesis.

Elimination of objects using learning algorithms solves problems more effectively than human eyesight. Convoluted neural networks are widely used in problems of classification, detection and recognition of images. Gradually, the range of these tasks is expanding, so the development of new architectures, layers of the network and modifications of software platforms is not lost.

Class is a certain group of objects in classification system that combines a certain set of objects by a certain feature or features. Perceiving the outside world, we always classify information, that is, we break them into groups of similar but not identical phenomena. For example, despite a significant difference, one group includes all the letters “A”, written in different handwritings or all sounds corresponding to the same note taken in any octave and on any instrument. To compose the concept of the perception group, it is sufficient to get acquainted with an insignificant number of its representatives [4].

II. THE PROBLEMFORMULATION

The specificity of the task that solves the object recognition system determines a number of requirements for the hardware platform. The platform should be a computer with sufficient processing power for fast image processing, as well as the available ability to connect additional devices, such as a video camera, analog and digital sensors. The platform must also have enough permanent memory to store all the necessary data. In addition, the ability to create a single network for integration with the Industrial Internet of Things, including data transmission, to gain access to the Internet to communicate with the server or use alternative channels for transferring information to a central computer node. If there are several objects of different categories on the image, each of them is recognized [5].

III. INDUSTRIAL INTERNET OF THINGS(IIoT)

Industrial Internet of Things is a system of combined computer networks and connected physical objects (things) with built-in sensors and software for data collection and exchange, with the possibility of remote monitoring and control in an automated mode, without human intervention.

Industrial IoT systems are aimed at minimal user intervention in the management and, accordingly, in independent management of technical processes with computer equipment.

There are features of the industrial Internet of things:

- all devices are controlled automatically from a single command center;
- high-frequency support of user actions;
- automated process of monitoring and managing the life cycle of equipment.

The introduction of computer vision systems into the industrial Internet of things helps to ensure control over compliance with safety requirements at work through automatic control and data analysis.

Based on the above, the task of object recognition is also relevant in the implementation of relevant developments in the IIoT system [6].

IV. IMAGERECOGNIZINGMETHODS

There are various algorithms that allow you to recognize images. The algorithm for learning the machine for pattern recognition, based on the method of secant hyperplanes, consists in approximating the parts of hyperplanes separating the hypersurface and consists of the following main stages:

- training (formation of a separating surface);
- secant planes;
- elimination of excess planes;
- elimination of excess parts of planes;
- recognition of new objects.

When using the method of parallel variants, several devices are trained simultaneously and independently of one another on the same material. When recognizing new objects, each machine will refer these objects to some image, maybe not to the same image. The final decision is made by "voting" machines, the object refers to the image to which it was attributed by a greater number of machines. The way to increase the reliability of recognition is to some improvement in the method of secant planes. It can be assumed that if the secant planes are drawn close to the plane passing through the middle of the straight line connecting the object and the opponent perpendicular to this line, the resultant surface will be closer to the true boundary between the images.

V. OBJECT RECOGNITION STAGES

The task of recognition of objects (images) is determined by the following steps:

- the definition of boundaries is the lowest-level task for which the neural networks are already classically applied;
- the definition of the vector to the normal allows us to reconstruct a three-dimensional image from a two-dimensional image;
- the definition of objects of attention (saliency) is what the person would pay maximum attention to when analyzing the picture;
- semantic segmentation allows you to divide objects into classes according to their structure, not knowing anything about these objects, that is, even before they are recognized;
- the semantic delineation of boundaries is the allocation of boundaries divided into classes;
- the highest-level task is the recognition of objects themselves [7].

VI. YOLO V3 AND V2 CHARACTERISTICS

Yolo v3 (and v2) is a program neural network packet aimed at object recognition. With its help, it is possible to localize and identify an object [8]. All previous detection systems use classifiers or localizers to perform tasks. They apply the model to the image in several places and scales. Areas with high detection density are considered to be detected [9].

Yolo uses a different approach. A single neural network is applied to the full image. This network divides the image into regions and predicts the limiting fields and probabilities for each region. These limiting fields are weighted by the projected probabilities.

The abbreviation YOLO stands for You Only Look Once. This model imposes a grid on the image, dividing it into cells. Each cell attempts to predict the coordinates of the detection zone with a confidence estimate for these fields and the probability of the classes. Then, the confidence estimate for each detection zone is multiplied by the probability of the class to obtain a final rating [10].

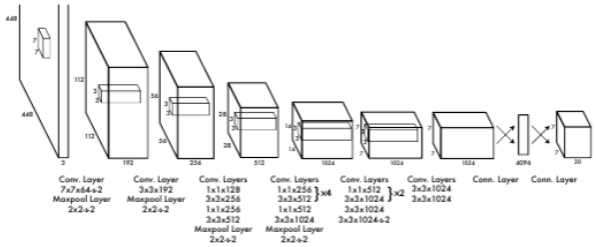

Fig. 1. The DarkNet YOLO v3 neural network architecture

Images that need to be processed can be obtained from different sources: a USB camera or .jpeg files placed in the / darknet / data directory. It is also possible to process video files.

The Yolo model was developed for a neural network based on DarkNet. DarkNet stores the learned coefficients (weights) in a format that can be recognized by different methods on different platforms. DarkNet is written in C and has no other programming interface, so if the platform requirements or your own preferences require you to access another programming language, you will have to work on its integration in addition. Also, it is distributed only in the source code format, and the compilation process on some platforms can be very problematic [11].

The neural network consists of 24 convolution layers, followed by 2 completely connected layers. The rotation of the 1 × 1 convolution layers separates the load on the previous layers. The resolution of the convolution layers in the problem of classifying images is halved (input image 224 × 224), but then the resolution for detection is doubled again [12].

In general, the task of object recognition consists of two parts: learning and recognition. Training is carried out by showing individual objects with indication of their belonging to one or another image. As a result of training, the recognition system should acquire the ability to react with identical reactions to all objects of the same image and other reactions to all objects of distinct images. It is very important that the learning process should be completed only by showing a finite number of objects. Learning objects can be either pictures or other visual images (letters, numbers). It is important that in the learning process only the objects and their belonging to the image are indicated. The training follows the process of recognition of new objects, which characterizes the actions of the already trained system. Automation of these procedures is the problem of learning to recognize patterns. In the case when a person himself unravels or thinks up, and then imposes a classification rule on the machine, the problem of recognition is partially solved, since the main and the main part of the task (training) is taken by the individual.

The input image falls into a network of layers, called filters of different sizes and different complexity of the elements that they recognize. These filters constitute a random index or a set

of characteristics, which then falls into the classifier. Usually it is a multilayer perceptron.

In the image and likeness with a biological neural network, objects of varying complexity are recognized.

VII. PROGRAM TESTING

Configuration of the personal computer (1):
CPU: Intel® Core ™ i7-2670QM CPU @ 2.20 GHz
Memory (RAM) 6.00 GB
Windows 10 Pro x64, Ubuntu 16.04 x64
Graphicscard: ATI Radeon HD6770M
Harddrive: 1.00TB

You need to install the OpenCV package. Installation of CUDA, cuDNN is not required, because on this device there is no video adapter of the corresponding manufacturer (NVidia) [13].

Install the package Darknet Yolo:

We put the file yolov3.weights in the directory darknet;

open the Makefile, set OpenCV and OpenMP = 1.

Next, we execute the assembly commands without leaving the directory.

Run the command for recognition from the web-camera:

./darknet detector demo cfg / coco.data cfg / yolov3.cfg yolov3.weights

The image will be received via USB-web-camera.

It is noticeable that in the terminal the frame processing frequency is counted, 0.2 FPS; it means that one frame leaves 1 / 0.2 = 5 seconds.

Also, the program calculates the percentage of authenticity of object recognition.

Let's estimate the speed of the system: put the image file WP_20160805_009.png in the directory / darknet / data.

Image processing results WP_20160805_009.png:

processing time = 4.85 s.

Repeat the same with another image (WP_20160819_001) .png:

image processing results WP_20160819_001.png:

processing time = 4.94 s.

It turns out that, on average, the image processing takes about 5 seconds.

VIII. PROGRAM TESTING (SECOND EQUIPMENT OPTION)

Let's take a computer of a different configuration (2):

CPU: AMD Athlon 2 x 3 440 @ 3 GHz

Memory (RAM) 4.00 GB

Windows 10 Pro x64, Ubuntu 16.04 x64

Graphics card: NVidia GeForce GTX650Ti 1 GB,

Harddrive: 380 GB

In the Makefile, set the GPU and CUDNN = 1.

For technical reasons, we launch not YOLO v3, but YOLO v2 (the terminal displays a message that there is not enough video memory for the YOLO v3 on the graphics processor, apparently, a higher-performance graphics card is required).

Fig. 2. WP_20160819_001.jpg image processing

We start the recognition from the camera. Immediately note the frequency of 8.9 frames per second. This is 45 times higher than the previous equipment (0.2 frames per second).

We will recognize the same images:

- image WP20160805_009.jpg is detected in 0.096 seconds;
- image of WP20160819_001.jpg was detected in 0.098 seconds.

On average, the processing of one image takes 0.1 s, which is 50 times faster than in the previous configuration (5 s). These examples do not show a backlog in the classification of objects by different versions of the program.

IX. IIoT INTEGRATION

It is considered permissible integration of the software and hardware complex into the IIoT service, namely, the delivery of recognition results to where other services can access them.

To date, a widely distributed solution is python daemon.py, which will run a simple server that displays a video stream from a webcam with forecasts for: http://127.0.0.1:8000/events/.

The task is relevant for monitoring the actions of personnel in production, as well as for monitoring the position of materials on the conveyor.

For the security of the system, data is encrypted. Different solutions are being developed, one of them is the automatic generation of the key by the client device.

Fig. 3. WP_20160805_009.jpg image processing

X. CONCLUSION

The system Yolo v3 and Yolo v2 has been tested. The architecture and process of image processing are considered. The system recognizes objects and defines their names. The neural network is tested on its own equipment.

It is noted that the use of NVidia GTX650Ti for processing graphics allows to process video stream from webcam 45 times faster, and images 50 times faster than on Intel Core i7 2.20 GHz second generation. The system operation is expedient when using the NVidia graphics processor with CUDA cores.

All detected objects are recognized, and the accuracy of the detection is calculated. The algorithms of working with the network are shown.

TABLE 1. PERFORMANCE COMPARISON ON DIFFERENT EQUIPMENT

№	Operation name	Run time (frame rate)		Performance factor (2 relative to 1)
		1	2	
1	Web-camera recognition	0.2 frames/s	8.9 frames/s	44.5 ≈ 45
2	WP_20160819_001 image recognition	4.94 s	0.098 s	50
3	WP_20160805_009 image recognition	4.85 s	0.096 s	50

XI. REFERENCES

[1] Воронов В.И., Усачев В.А. Компетенция «машинное обучение и большие данные» // Приоритетные направления развития науки и образования / Под общ. ред. Г.Ю. Гуляева. Пенза, 2017. С. 97–108. [Usachev V.A., Voronov V.I. The competence “Machine learning and Big Data” // Priorities for development of science and education / Ed. G.Yu. Gulayev. Pensa, 2017. P. 97–108].

[2] Михаеску С.В., Трунов А.С., Воронова Л.И. Анализ предметной области для разработки системы построения скелетной модели человека на основе массива опорных точек, получаемых совокупностью контроллеров Kinect // Международный студенческий научный вестник. 2015. № 3-4. С. 521–522. [Mihaesku S.V., Trunov A.S., Voronova L.I. Analysis of the subject area of system developing for building a human skeletal model based on an array of reference points obtained from a set of Kinect controllers // International Student Scientific Bulletin, 2015. N 3-4. P. 521–522].

[3] Воронова Л.И., Воронов В.И. Machine Learning: Регрессионные методы интеллектуального анализа данных: учеб. пос. / МТУСИ. М., 2017. [Voronova L. I., Voronov V. I.. Machine Learning: Regression methods data mining: A tutorial / MTUCI. Moscow, 2017].

[4] Machine Learning. Seminars on neural networks. URL: http://www.machinelearning.ru/wiki/images/1/1e/Sem07_ann.pdf.

[5] Молодяков С.А., Тышкевич А.И. Принципы выделения параллельных потоков команд обработки видеоизображенгий в smart-видеокамерах // Международный научный журнал. 2016. № 9. С. 76–80. [Molodyakov S.A., Tishkevich A. I. Principles of allocation of parallel streams of commands of processing of video images in smart-video cameras // International science journal. 2016. N 9. P. 76–80].

[6] Consumer Home // Digital Living Network Alliance. URL: http://www.dlna.org.

[7] Reference Model for Service Oriented Architecture // Organization for the Advancement of Structured Information Standards. URL: http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf.

[8] Redmon J., Divvala S., Girshick R. et al. You Only Look Once: Unified, Real-Time Object Detection // The conputer vision Foundation. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.

[9] Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. 2-е изд., стер. М.: Горячая линия — Телеком, 2002.[Kruglov V. V., Borisov V. V. Artificial neural networks. Theory and practice. 2nd ed., ster. Moscow: Goryachaya Liniya — Telecom, 2002].

[10] Agrawal P., Girshick R., Malik J. Analyzing the performance of multilayer neural networks for object recognition // ECCV. 2014. P. 329–344.

[11] Bengio Y., Courville A.C., Vincent P. Unsupervised feature learning and deep learning: A review and new perspectives // CoRR. 2012. Vol. 1. abs/1206.5538.

[12] Redmon J., Angelova A.. Real-time grasp detection usingconvolutional neural networks // CoRR. 2014. abs/1412.3128.

[13] Буянов Б.Я., Верба В.А. Использование модуля dnn библиотеки OpenCV 3.3 для распознавания объектов // Инновационные технологии в кинематографе и образовании: Сб. IV Междунар. науч.-практ. конф. / ВГИК. М., 2017. С. 47–56. [Buyanov B.Ya., Verba V.A. Using the dnn module of the OpenCV 3.3 library for object recognition // Innovative technologies in cinema and education: Coll. IV International scientific-practical conf. / VGIK. Moscow, 2017. P. 47–56].