УДК 004.8

Maksym Mazur

PROBLEM OF DETECTION AND CLASSIFICATION OF OBJECTS WITH CONVOLUTIONAL NEURAL NETWORKS

Vinnytsa National Technical University

Abstract

The principle of work of convolutional neural networks, methodical identification and classification of objects using stellar neural networks, and their features are considered.

Keywords: Convolutional neural network, CNN, classification, detection, network model.

Computer vision is an interdisciplinary field, which has gained a lot of attention in recent years (since CNN), especially cars with autopilot, that that were in a spotlight. Another major part of computer vision is the detection of objects. Instruments of detecting objects according to the position on the image, detection of vehicles, video surveillance, etc. The difference between object detection algorithms and classification algorithms is that in the detection algorithms we try to draw a bounding frame around an object of interest to find it in the image. In addition, it is not necessary to draw only one restriction block in the event of object discovery, there may be many restrictive blocks representing various objects of interest in the image, and it is not known how many they may be in one image [1].

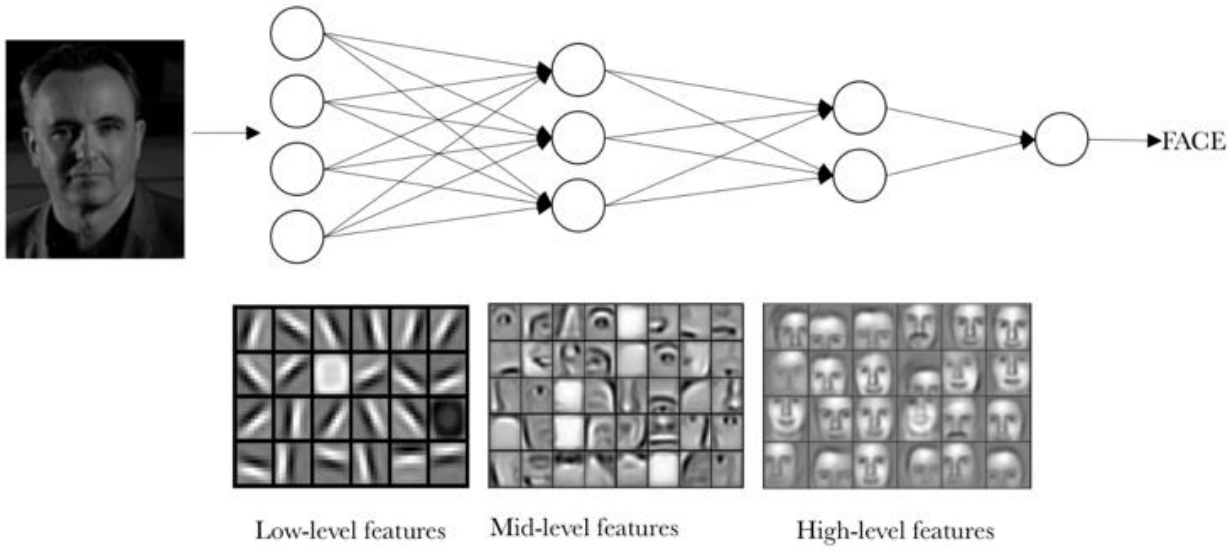
Convolutional neural network (CNN) is a specific casual neural network with deep learning that was already in use in the late 90's but has become extremely popular in recent years when it became the part of computer vision development.

Convolutional neural networks have many common features with other neural networks: they are formed by neurons that have parameters in the form of weights and bias that can be learnt. However, main feature of CNN is that they are able to handle images well in the form of data arrays, which allows us to encode certain properties in architecture for the recognition of specific objects in images [2].

To get an intuitive idea of how these neural networks work, it's worth paying attention to how people recognize things. For example, if we see a person, we recognize personality because he or she has his or her ears, eyes, nose, hair, and so on. Then, to decide who is in front of us, for example, we check some imaginary checking areas that we take into account. Sometimes you can`t see the human ear, because it is closed to the hair, but we also classify the face with a certain probability, because we see the eyes, nose and mouth. Actually, it is the principle that classifiers based on convolutional neural networks has in basis. But in fact, we must first know what an ear or a nose is to see if they are in the image; this means that we must earlier identify the lines, borders, contours or shapes that are similar to those that contain the ear or nose that we saw earlier. Convolutional neural networks deal with this part.

To know that some of these parts of the human face are on the picture is not enough to say that this is human. We must also be able to determine how parts of the face are situated with each other, their relative sizes, etc.; otherwise, the face will not look as we are used to. Visually, it can be imagined that the first layers of the convolution network contain a set of contours and forms, which in the subsequent layers bind the individual parts of the image and search or match the location of individual parts of the desired image (pic. 1). Each layer is responsible for the level of abstraction.

But the problem of detection and classification can`t be solved by constructing a standard convolutional network followed by a fully connected layer, since the length of the output layer is variable - not constant because the number of objects that can enter the image is not fixed. The naive approach to solving this problem is to take from the image various areas of interest, and use the ZNM to classify the presence of an object within the scope. The problem with this approach is that objects of interest may have different spatial locations within the image and different aspect ratios. So, you have to choose a huge number of regions, and it can use a lot of computing power. Thus, algorithms such as R-CNN, YOLO, etc., have been developed to find a reduction in the number of computations and to increase accuracy.



Picture 1 - Visual representation of the first layers of the convolutional neural network

R-CNN

To overcome the problem of choosing a huge number of areas, Ros Hirschik et al. proposed a method in which we use a selective search to extract only 2000 domains from the image, and he called them regional interest zones [3]. Therefore, now, instead of classifying a huge number of regions, you can work with 2000 regions. These 2000 regional proposals are generated using the selective search algorithm:

1. Create an initial subsegmentation, generate a lot of candidate regions;
2. Use the greedy algorithm for recursive combining of such regions into large ones;
3. Use the generated regions to create the final proposals of the candidate region.

These 2000 proposals from candidate regions are distorted in the square and come into the convolution of the neural network, which produces a 4096-dimensional vector of signs as a source. NNM acts as an extractor of signs and the output dense layer consists of functions extracted from the image, and the elongated attributes are fed into the SVM to classify the presence of an object in this proposal of the candidate region. In addition to predicting the presence of an object within the region's offerings. The algorithm also provides four values that are biasing values for increasing the accuracy of the restrictive region. For example, given the region's offer, the algorithm would involve a person's presence, but the person's face within this region's offer could be halved.

The main problems of the R-CNN algorithm are:

-There is plenty of time to train the network, since it is necessary to classify 2000 regions for each image.

-The method can`t be implemented in real time, as it takes about 47 seconds for each test image.

-The algorithm for selective search is a fixed algorithm. Therefore, there is no training at this stage. This may lead to poor proposals from the candidate regions.

Fast R-CNN

The author of previous work (R-CNN) has solved some of the disadvantages of R-CNN to build a faster object detection algorithm, called the Fast R-CNN. The approach is similar to the R-CNN algorithm. However, instead of submitting regional offers to CNN, we submit the input to the network to create a rollback property map. On a map of convolutional characteristics, we identify the fields of proposals and deform them in squares and using the union layer (RoI), we change them to a fixed size so that it can enter into a fully connected layer (Fully connected layer). We use the softmax layer to predict the class of the proposed region, as well as the offset values for the restrictive region from the RoI limitation vector.

The reason that the Fast R-CNN is faster than R-CNN is that you do not have to submit the region's offerings to convolutional neural networks every time. Instead, the convolution operation is executed only once per image, and a map of objects is generated from it.

Faster R-CNN

Both of the above-mentioned algorithms (R-CNN & Fast R-CNN) use a selective search to find region offers. Selective search is a slow and time-consuming process that affects network performance. Therefore, Shaoqing Ren et al. has invented an algorithm for detecting objects that excludes the sampling algorithm and allows the network to explore the region's proposals.

Like the fast R-CNN, the image is provided as an input signal for a convolutional network that provides a convolution. Instead of using the selective search algorithm on a map of objects to identify the region's proposals, a separate network is used to predict the region's proposals. The predicted areas of the proposals are then modified using the RoI pool layer, which is then used to classify the image within the proposed region and predict the offset values for the bounding fields.

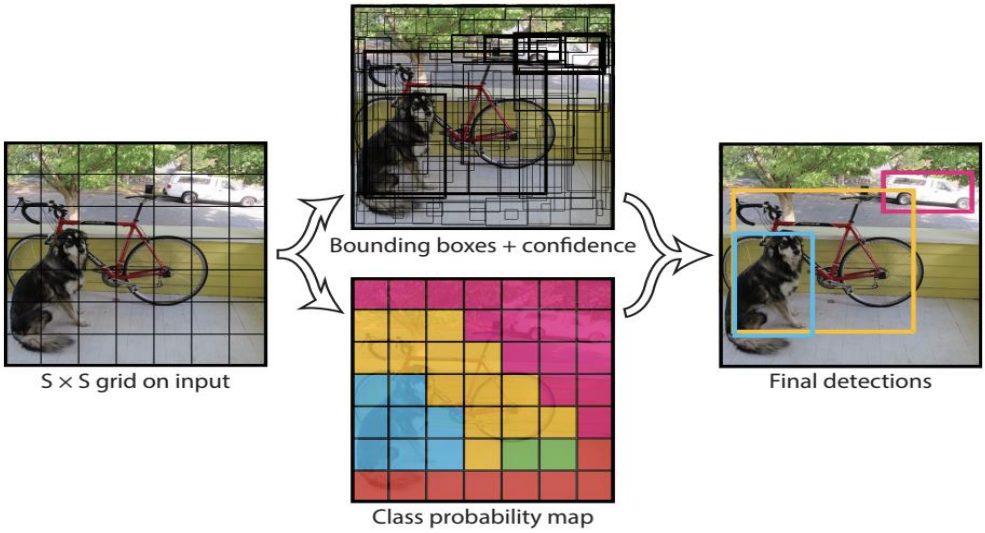
YOLO(You Only Look Once)

All previous detection algorithms use objects to localize an object in an image. The network does not look at the full image. Instead, parts of the image that have high probabilities contain an object. YOLO or You Only Look Once is an algorithm for detecting objects that is significantly different from algorithms based on the selection of random regions. In YOLO, a single roller network provides a limiting framework and a class probability for these regions [4].

The image is split into SxS grid, we take m bounding fields within each grid. For each bounding field, the network outputs the values of class probabilities and offsets for the bounding field. Limiting frames that have a class probability above a threshold value are selected and used to find an object in an image.

YOLO is an order of magnitude faster (45 frames per second) than other algorithms for detecting objects. The limitation of the YOLO algorithm is that it can`t handle small objects in the image, for example, it may have difficulty detecting a flock of birds. This is due to the spatial constraints of the algorithm.

Conferences on computer vision every year are revising new radical concepts, and each year roller networks are increasingly being used to solve computer vision problems.



Picture 2 - Visualization of the partition of the image into zones by the YOLO method

REFERENCES

1. Колесницький О.К. Моделювання імпульсної нейронної мережі у задачі розпізнавання багатовимірних імпульсних послідовностей / О.К. Колесницький, С.М. Богатчук, М.В.Крещенецька, С.С.Яремчук // Вісник Вінницького політехнічного інституту. — Вінниця: УНІВЕРСУМ-Вінниця, 2008. — №5, С.62-66.
2. Convolutional Neural Networks — [Electronic resource]. — Access mode: https://towardsdatascience.com/convolutional-neural-networks-for-beginners-practical-guide-with-python-and-keras-dc688ea90dca
3. R-CNN, Fast R-CNN, Faster R-CNN, YOLO—Object Detection Algorithms — [[Electronic resource]. — Access mode: https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e
4. YOLO: Real-Time Object Detection — [[Electronic resource]. — Access mode: https://pjreddie.com/darknet/yolo/.

Maksym Mazur — student, group 1CS-18m, Faculty of information technology and computer engineering, Vinnytsa National Technical University, Vinnytsia.

Науковий керівник:
Присяжна Олеся Дмитрівна кандидат філологічних наук, старший викладач кафедри іноземних мов Вінницького національного технічного університету prysyazhnalesya@gmail.com

Prysyazhna Olesya Dmitrievna Candidate of Philology Sciences, Senior Lecturer of English, the Foreign Languages Department, Vinnytsia National Technical University, Vinnytsia prysyazhnalesya@gmail.com