On The Analysis of Some Deep Learning Algorithms for Object Detection and Applications

Sandy Montajab Hazzouri
Faculty of Informatics Engineering, Albaath University, Syria
Samonhaco1994@gmail.com

Abstract

The traditional methods of discovering objects no longer meet the requirements of the times as a result of their reliance on non-dynamic methods and as a result of their slow performance in light of the world's dependence on a huge amount of multimedia and social media. With the rapid development of deep learning providing more powerful tools capable of manipulating high-level and complex semantic features of objects. Several techniques have been developed to detect objects using deep learning algorithms. This research presents a comparative analysis of the most famous deep learning techniques for object detection, explaining their mechanisms, use cases and an experimental evaluation of their performance.

Keywords: Algorithm; Deep learning; Object detection; Model

1. Introduction

CNN deep networks significantly improve the classification of images and increase the accuracy of object detection. The process of detecting objects is more difficult compared to classifying images, so it requires more complex methods to solve it, and this complexity arises because the detection process requires accurate positioning of objects, so traditional methods of training models are usually slow and inaccurate. Which creates two main challenges:
1. Several locations of candidate objects must be processed (they are often called "Proposals").
2. Only an approximate location can be made and then it must be adjusted to achieve a more accurate determination.
The solutions to these challenges often affect the speed, accuracy or simplicity of the design.
In this research, we will compare the most commonly used techniques for detecting objects.

2. Research objective

This research aims to provide a comparative analysis of the most popular object detection technologies using deep learning. Provide a simplified explanation of each technique with its use cases, strengths and mechanism of action, and then conduct experiments to measure and compare their performance.

3. Previous studies

- The **research paper [1]** discusses the possibility of using a pre-designed CNN network to classify images and add the task of detecting objects to it. The paper concludes that it is possible to develop a simple system that enhances the average accuracy by more than 40%. The framework combines powerful computer vision techniques for generating bottom-up region propositions with recent advances in learning .CNN the system operates as a set of experiments that provide insight into what the network is learning to represent, revealing a rich hierarchy of discriminating and often meaningful features.

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

- The **research paper [2]** presents a different way to do object class discovery using CNN networks, which is much faster than the latest alternative solutions. Many modern object detection systems are based on .object proposals this paper proposes to divide the image into grid cells based on the study of the probability of the presence of objects in the cells as well as the study of the slope of the bounding box of each object. Speed is the main advantage because the network processes the image only once and detects objects. The results are very good. The accuracy is lower than other techniques by about 7%, but it is much faster at 45 FPS inference.
- The **research paper [3]** provides a comprehensive survey of recent developments in the detection of visual objects using deep learning. By reviewing a large body of relevant recent work, the researchers systematically analyzed the existing object detection frameworks and organized the survey into three main parts: (1) detection components, (2) Learning Strategies, and (3) applications and standards. In the survey, we cover a variety of factors that affect detection performance in detail, such as detector architectures, feature learning, proposal creation, sampling strategies, etc. Finally, the researchers discuss several future directions to facilitate and stimulate future research to discover visual objects with deep learning.
- The **research paper [4]** provides a comprehensive review of research related to the discovery of video objects. This research attempts to link and systematize the latest cutting-edge research on object detection in video with the aim of classifying and loading the algorithms used based on specific representative models. The differences and connections between object detection in video and similar tasks are systematically presented, and evaluation and performance measures for object detection in video are presented for approximately 40 models in two datasets. Finally, the various applications and challenges of object detection are discussed in the video.

4. Deep Learning

Before talking about object detection technologies based on deep learning, we will give a review about deep learning along with an introduction about the infrastructure and advantages of CNN.

4.1. Introduction to deep learning

Deep models can be referred to as neural networks with a deep structure, that is, they include a larger number of hidden layers. The history of neural networks can be traced back to the Forties of the last century [5]. Deep learning became popular a year ago. [6] 2006 the prevalence of deep learning can be attributed to the following factors:

- The appearance of a number of parameter training data (annotated), such as ImageNet [7].
- Development of high-performance computing systems, such as gpuu and TPU clusters.
- Significant developments in the design of network structures and training strategies. So that it becomes

Training of deep neural networks is very effective. Meanwhile, various network architectures, such as AlexNet [8], overfit [9], GoogLeNet [10], VGG [11] and ResNet [12], have been extensively studied to improve performance.

4.2. CNN structure and features

A convolutional neural network (also known as ConvNet, or CNN) is a type of front-Fed neural network used for tasks such as image loading, natural language processing, and other complex image classification problems [13]. It is distinguished in that it can pick out and detect patterns from images and texts and understand them.

Convolutional neural networks look at one patch of the image at a time and gradually scan the image in this way to extract complete information [8].

4.2.1. CNN layers [13]

- Convolutional Layer
- RELU layer
- Pooling Layer
- Normalization Layer
- Fully-Connected Layer

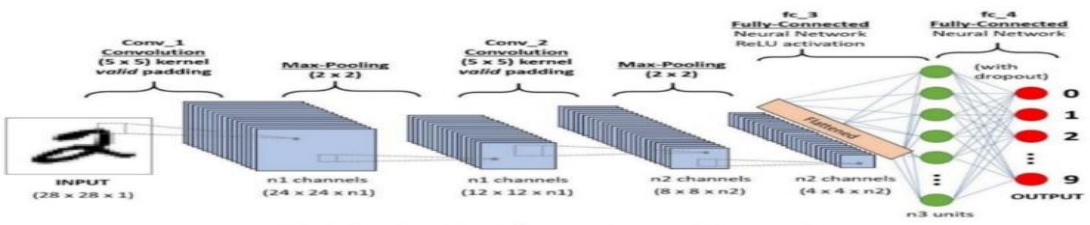

Figure 1. CNN sequence for classifying handwritten numbers [8]

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

5. Object detection techniques using deep learning

Object detection is a computer vision technology that allows us to identify what and locate objects in an image or video. Object detection can be used to count the number of objects in a scene, pinpoint their exact locations, and track them, all while accurately tagging them.

For example, a picture with two cats and a person. Object detection allows us to classify the types of found objects with the location of their occurrence determined in the image.

Figure 2. An example of object detection

5.1. What is Object Detection?

Object detection is a computer vision technique that serves to recognize and locate objects within an image or video. Specifically, object detection draws bounding boxes around these detected objects, which allows us to determine where said objects are located (or how they move) in a given scene.

5.2. The difference between image classification and object detection:

Image classification and object detection techniques are important methods when it comes to working in computer vision. These technologies help machines understand and identify objects and environments in real time with the help of digital images as input.

5.2.1. Image Classification

Image classification is a technique used to classify or predict the category of a particular object in an image. The main goal of this technique is to accurately determine the features in the image. In general, image classification techniques can be classified as (parametric and non-parametric) or (supervised and unsupervised). For supervised classification, this technique provides results based on the established decision limits, which mostly depend on the inputs and outputs provided during the training of the model. But in the case of unsupervised classification, the technique provides the result based on the analysis of its input dataset; features are not fed directly to the models.

The main steps involved in image classification techniques are in .1determine a suitable classification system .2. Extraction features .3. Choose good training samples .4. Image pre-processing .5. Choose the appropriate classification method .6and processing of the subsequent classification, 7. Finally evaluate the overall accuracy. In this technique, the inputs are usually an image of a specific object, and the outputs are the expected categories that identify and match the input objects. Convolutional neural networks (CNNs) are the most common neural network models used in the image classification problem. [14].

5.2.2. Discovery of objects

The definition of the object detection problem is to determine where the objects are located in a particular image and to which category each object belongs. In simple words, object detection is a kind of image classification technique, and besides classification, this technique also determines the location of the object.

This technology has the ability to search for a certain category of objects, such as cars, people, animals, birds, etc., and has been successfully used in video processing systems. Recent developments in this technique have become possible only with the advent of deep learning.

The steps of detecting conventional objects can be mainly divided into three stages, namely, the selection of the information-containing area, feature extraction, and classification.

5.3. Stations develop object detection

The field of discovering things is not as new as it might seem. In fact, the discovery of objects has evolved over the past twenty years. The progress of the discovery of objects is usually separated into two separate historical periods (before and after the introduction of deep learning):

Before 2014- traditional methods before deep learning

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

- [15] (2001) Viola-Jones Detector, the pioneering work that initiated the development of methods for detecting objects
- [16] (2006) HOG Detector, describing a common feature of object detection in computer vision and image processing.
- DPM (2008), with the first introduction to the surrounding box Regression.

After 2014-methods based on deep learning.
The most important two-stage object detection algorithms

- [1] (2014)RCNN and SPPNet
- [18] [17] (2015) Fast RCNN and Faster RCNN
- [19] (2017) Mask R-CNN
- [24] (2017) Pyramid Networks/FPN
- [21] (2021) G-RCNN

The most important single-stage object detection algorithms

- [2] (2016) YOLO
- [22] (2016) SSD
- (2017) RetinaNet [23]
- YOLOv3 (2018) [24]
- YOLOv4 (2020) [25]
- YOLOR (2021)

As we can see in the list above, modern object detection methods can be classified into two main types: single-stage object detection technologies versus two-stage object detection technologies.
In general, object detection techniques based on deep learning extract features from the input image or video frame. The object detection algorithm solves two consecutive tasks:

- Task number: 1 Find a random number of objects
- Task number: 2classify each object separately and estimate its size by a bounding box.

To simplify the process, these tasks can be separated into two stages. While other methods combine both tasks in one step (single-stage detection techniques) to achieve higher performance at the expense of accuracy.

5.4. Detection of objects in two stages

In this section, we will delve into the main ideas of the stages of the work of multistage algorithms for detecting objects by reviewing some of the most important papers in this field.
One model is used to extract areas of objects, and a second one is used to classify and further improve the localization of the object. It is known that such methods are relatively slow, but very powerful, however recent advances such as feature sharing, have led to the optimization of two-stage detectors to obtain a similar computational cost with single-stage detection models.
These works are highly dependent on previous works and are mostly based on previous stages as infrastructure. Therefore, it is important to understand all the main algorithms in two-stage detection models.

5.4.1. R-CNN

A year [1] 2014 paper proposes a simple version of the CNN-based two-stage object detection algorithm that has been improved and accelerated in the following papers. As shown in Figure 4, the workflow consists of three stages:
1. Create Area proposals: the form should draw nominations for the objects in the picture, independently of the category (specify the area in which an object is expected to be located).
2. The second stage is a fully convolutional neural network that calculates features from each candidate region.
3. The final stage is a fully connected layer, which is expressed in the sheet as SVMs.

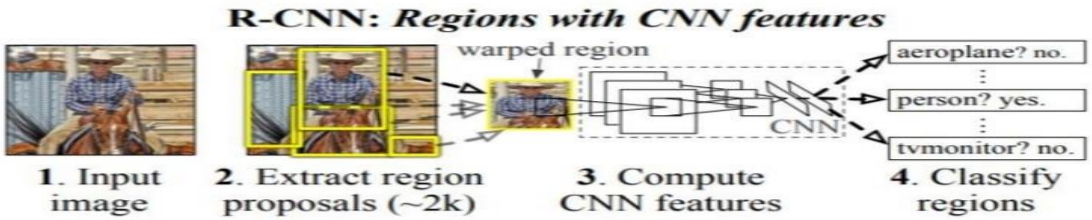

Figure 3. overview of the stages of R-CNN

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

Proposals for regions can be created using various methods, the paper chooses the use of selective search [26].
In a selective search, a hash algorithm is applied to the image, and area proposals (surrounding squares) are drawn based on the hash map. The hash map is repeatedly integrated (Bottom-Up) and the proposals for the larger area of the improved map are gradually drawn up as shown in Figure 5.

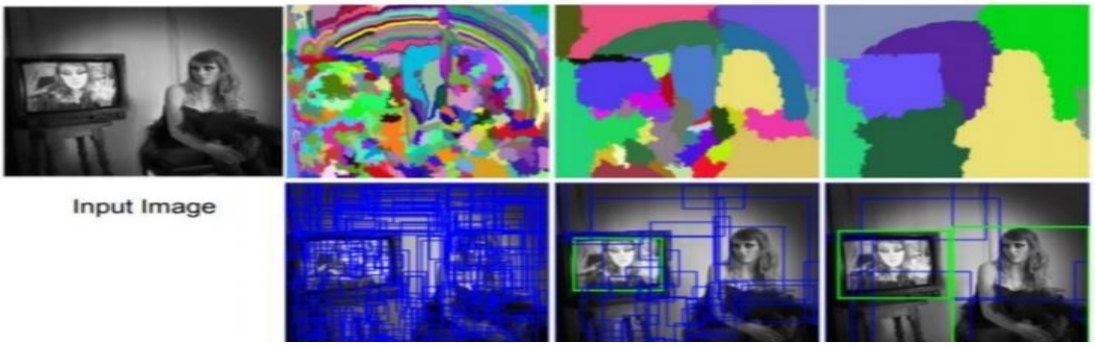

Figure 4. selective search

This stage (Selective Search) works as follows
1. First, similarities in color, structure, area and filling are used for non-object-based segmentation (segmentation). So we get several small fragmented areas as shown in the lower left of the image above.
2. Next, a Bottom-Up approach is used, in which small fragmented areas are combined together to form larger fragmented areas.
3. Thus, several proposals are being created for the areas of the bounding box (candidates) as shown in the photo.
After completing the selective search, we take each bounding box resulting from this mercy and enter it into the next stage, which is the stage of discovering the features of the potential object. But before that, a geometric correction must be made on this part of the image specified by the bounding box, because the second stage, which is the CNN network, accepts only a fixed income in terms of measurements, while the output of the selective search is area proposals of different sizes and therefore all possible areas are converted into squares. The paper [1] is used.
CNN is of the AlexNet style as a second stage, while any other CNN architecture can be used.

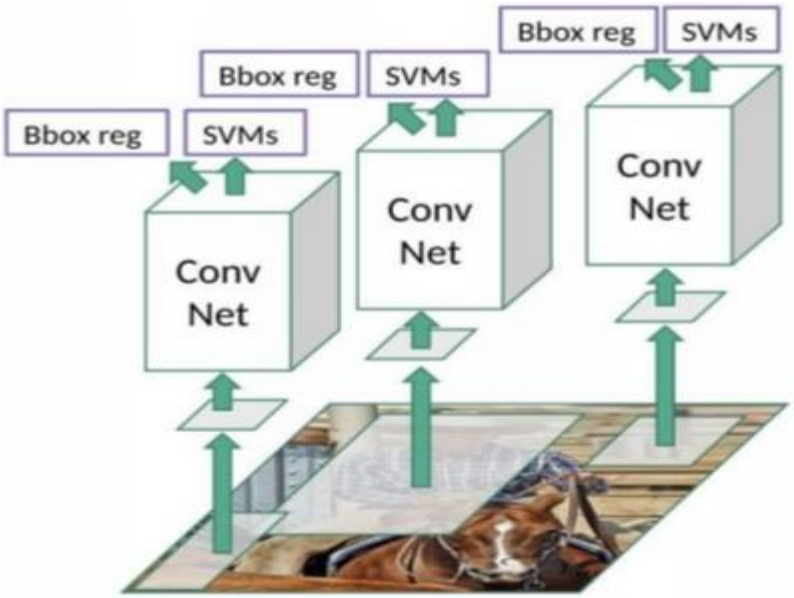

Figure 5. R-CNN operation

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

In the second stage, the proposed areas of selective search are fed individually to the convolutional layers of the CNN, in order to extract features.
The third stage is the object prediction stage and is done via feature processing by fully connected layers in .CNN
After the prediction process, the bounding box is corrected more accurately based on the prediction of the object. Where the authors also use (bounding box classifier) which has been trained to further improve the estimation of the bounding box made by hashing.
To reject the proposals of the overlapping area, where two or more surrounding squares point to the same object, the authors propose an algorithm that rejects the area if it has a high IoU (intersection-over-union) with another area with a more confident prediction.
The outperformance of R-CNN in comparison with other methods comes from the idea of performing selective bottom-up search also using CNN grids, to increase the accuracy of locating objects. This work combines computer vision, traditional image processing and deep learning to improve object detection.
But R-CNN's problems boil down to this:

- The process takes a long time to train the network as your client will have to classify 2000 zone suggestions for each image.
- It cannot be executed in real time as it takes about 47 seconds for each test image.
- The selective search algorithm is a static algorithm. Therefore, there is no learning happening at that stage. This may lead to the generation of bad proposals for the candidate region.

5.4.2. Fast R-CNN

The same author of the previous paper [1] solved some disadvantages of R-CNN to build a faster algorithm for detecting objects and named.[17] The Fast R-CN approach is similar to the algorithm .R-CNN however, instead of submitting area suggestions to CNN, We feed the input image to CNN to create a feature map. From the feature map, we determine the area of the proposals, correct it into squares and using a layer (RoI pooling), we reshape it to a fixed size so that it can be fed into a fully connected layer. From the beam of the RoI feature, we use the SoftMax layer to predict the proposed area class as well as the offset values of the surrounding square.

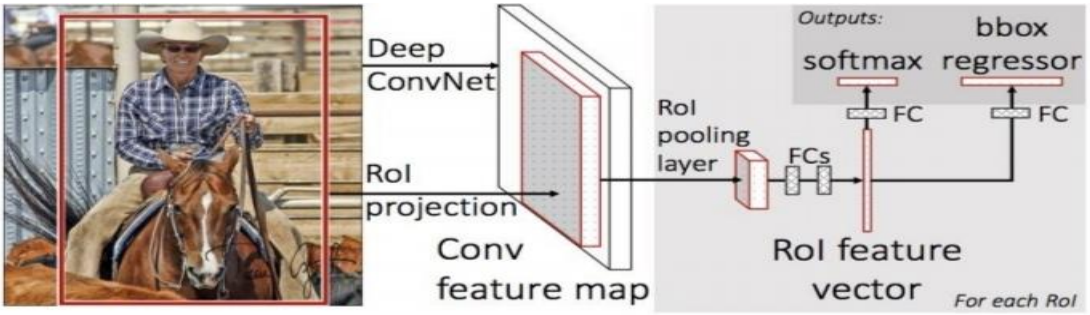

Figure 6. Fast R-CNN

The reason why Fast R-CNN is faster than R-CNN is that we don't have to submit 2000 zone suggestions to the convolutional neural network at a time. Instead, the wrapping process is done only once for each image and the landmark map is generated from it.
The following are the main contributions of the research paper [17]
1. A new layer proposal called ROI pooling extracts feature rays of equal length from all proposals (i.e. (ROIs) in the same image.
2. Compared to R-CNN, which has multiple stages (creating area proposals, feature extraction, and classification using SVM), Fast R-CNN builds a network that has only one stage.
3. R-CNN shares calculations (i.e. bypass layer calculations) across all proposals (RoI) rather than performing calculations for each proposal independently. This is done using the ROI pooling layer, which makes Fast R-CNN faster than R-CNN
4. Fast R-CNN does not cache extracted features, and therefore does not need much disk storage compared to R-CNN, which needs hundreds of gigabytes.
5. Makes Fast R-CNN more accurate than R-CNN.
Feeding the feature map from the last bypass layer to the ROI layer the reason for this is to extract a fixed-length feature vector from each area proposal.
Simply put, the ROI pooling layer works by dividing each zone proposal into a grid of cells. The max pooling operation

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

is applied to each cell in the grid to return one value. All values of all cells represent vector parameters. If the mesh size is 2×2, then the length of the feature vector is 4.

The extracted feature vector is then passed using ROI pooling to some fully connected layers .FC the output of the last FC layer is divided into 2 branches:

- Softmax layer for predicting class grades
- The FC layer is fully connected to predict the squares surrounding the detected objects.

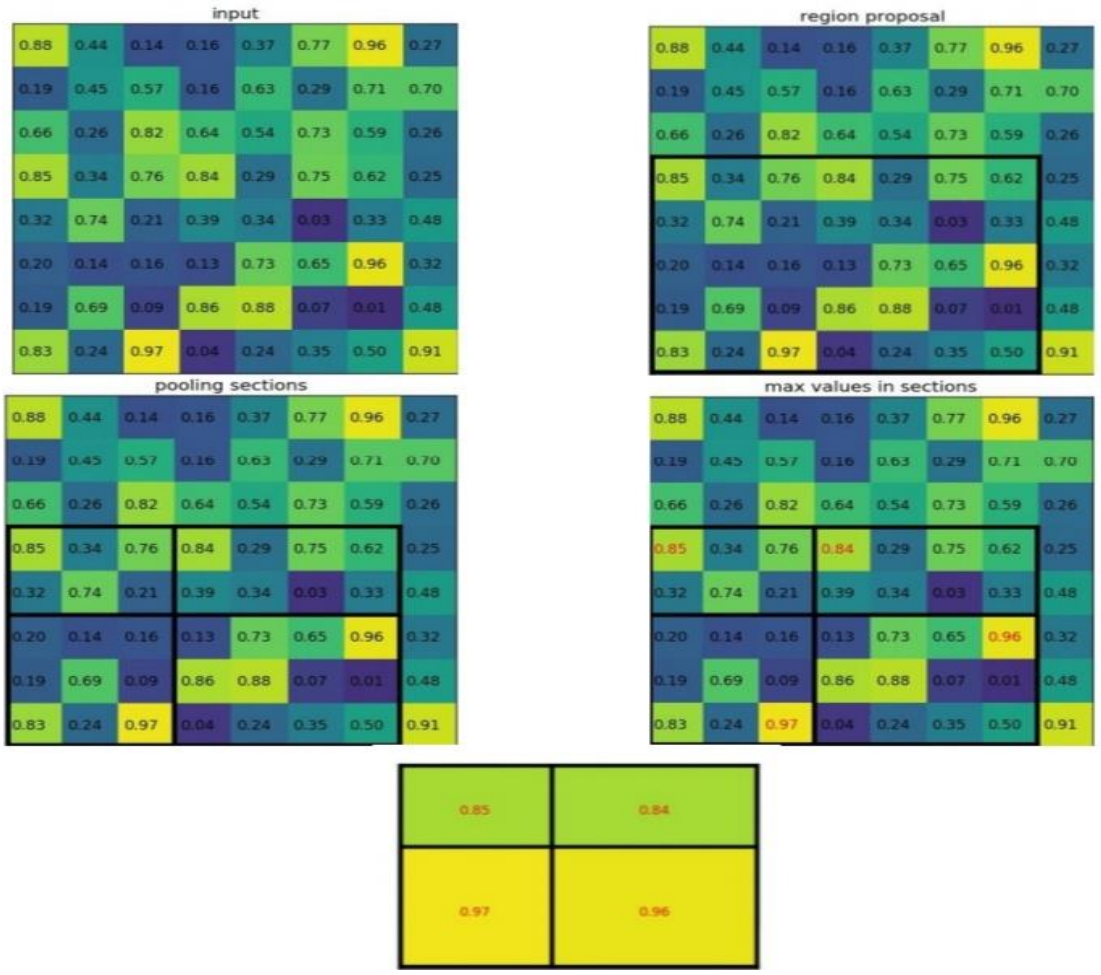

Figure 7. The process of determining the ROI

Despite the advantages of the Fast R-CNN model, there is a critical drawback because it relies on a selective search algorithm that takes a long time to create Area proposals. The selective search method cannot be customized on the task of discovering a specific object. Thus, it may not be accurate enough to detect all the target objects in the dataset.

5.4.3. Fast R-CNN

It was mentioned in the research paper [18]. It is an extension object detection algorithm for Fast R-CNN but it is faster than it thanks to the use of region proposal network (RPN) the main contributions of the paper [18].

1. The zone proposal network (RPN) is a fully convolutional network that generates proposals of different scales and width-to-height ratios. RPN implements the concept of a neural network with attention that tells Object Detection Technology (Fast R-CNN) where to look.

2. Instead of using image pyramids (i.e. multiple copies of the image but at different levels) or filter pyramids (i.e. multiple filters of different sizes,) this paper introduced the concept of Link boxes. The anchor box is a reference box of a certain scale and aspect ratio. With multiple reference anchor boxes, there are multiple scales and width-to-height ratios for the single zone. It can be considered that pyramid of reference anchor boxes each area is assigned to each reference anchor box, thereby detecting objects at different scales and aspect ratios.

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

5.4.4. Convolutional accounts are shared via RPN and .Fast R-CNN this reduces the computational time.
The structure of Faster R-CNN is shown in the following figure. It consists of 2 modules:

- RPN: to generate area proposals.
- Fast R-CNN: to detect objects in the proposed areas.

The RPN unit is responsible for generating area proposals. The concept of (attention) is applied in neural networks, so it is directed to the Fast R-CNN module where to search for objects in the image.
Notice how the output of convolutional layers is shared across both RPN and Fast R - CNN modules.

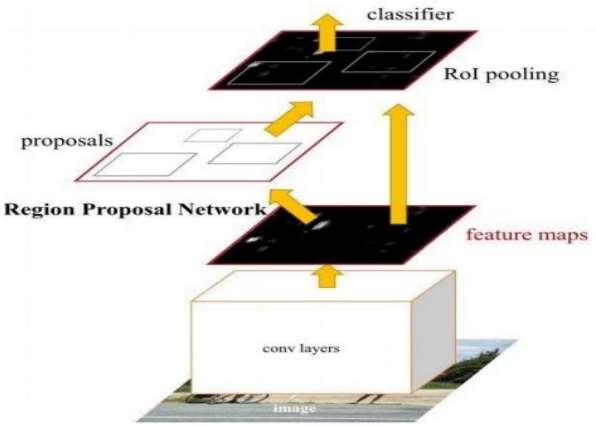

Figure 8. The Mechanism of action of Faster R-CNN

Faster R-CNN works as follows:
1. RPN is creating area proposals.
2. For all region propositions in the image, a fixed-length feature vector is extracted from each region using the ROI layer
3. The rays of the extracted features are then classified using. Faster R- CNN
4. The accuracy of the category prediction is returned for the detected objects as well as their surrounding squares.
The detailed mechanism of RPN work can be viewed by looking at the paper

5.5. Single-stage object detection:
We will focus on the model structure that directly predicts the squares surrounding the object for an image in a one-stage way. In other words, there is no intermediate task (as we discussed earlier in the region proposal) that must be performed in order to produce an output. This method leads to a simpler and faster model structure.

5.5.1. You Only Look Once (YOLO):
YOLO proposes to have a unified network to perform everything at once. Also, a comprehensive training network can be achieved [2].

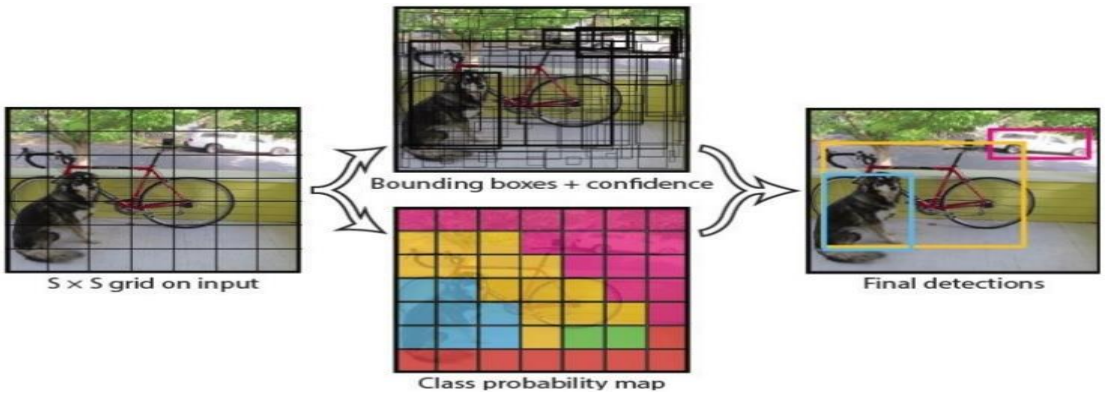

Figure 9. YOLO Unified Detection

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

The input image is divided into an $S \times S$ *grid* ($S = 7$). If the center of the object falls into a grid cell, then this grid cell is responsible for detecting this object.
Each grid cell expects enclosing squares B ($B = 2$) and confidence scores to enable the squares.
These confidence scores reflect how confident the model is that the box contains an object P
Each perimeter box consists of 5 predictions: *x, y, w, h, (confidence)*

- The coordinates (*x, y*), represent the center of the square relative to the boundaries of the grid cell.
- Width W and height h are projected for the whole image.
- The IOU trust between the expected square and any fact square represents a fundamental.

Each grid cell also predicts the probabilities of the conditional category,).P (Class | Object) total number of classes = 20).

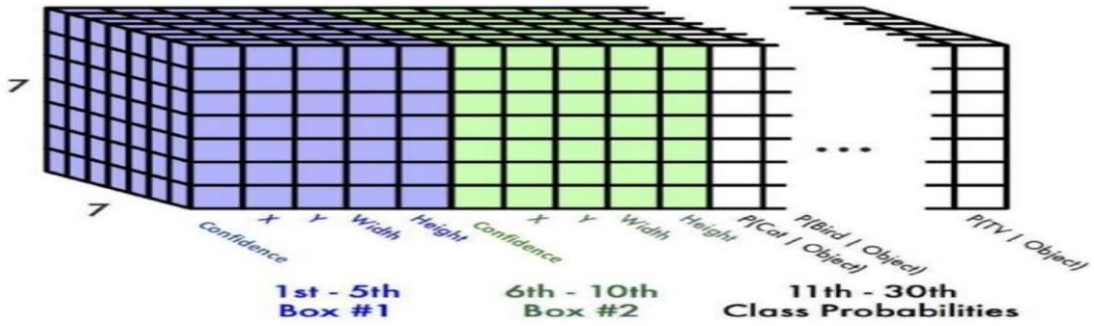

Figure 10. Yolo output

The network output is illustrated below:
The output volume becomes: $7 \times 7 \times (2 \times 5 + 20) = 1470$

5.5.1.1. Yolo network architecture:

The model consists of 24 convolutional layers followed by two fully connected ones. Alternating convolutional layers 1×1 work to reduce the area of features from the previous layers. (A 1×1 conversion was used in GoogLeNet to reduce the number of parameters).

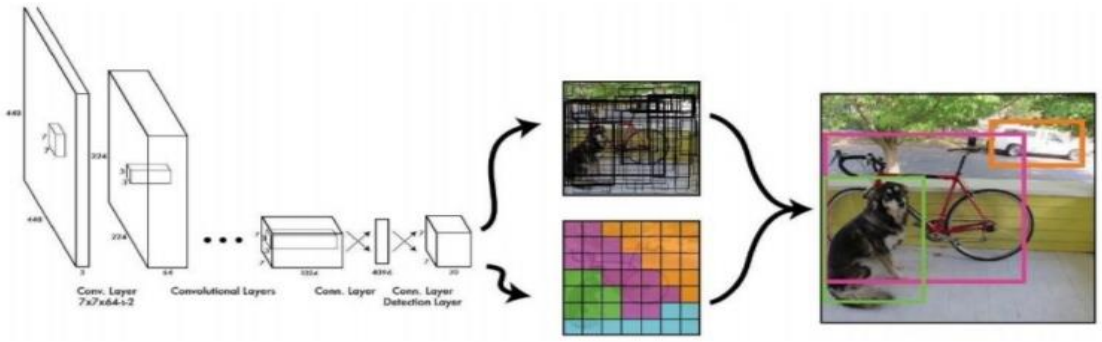

Figure 11. Stages of YOLO

Therefore, we can see that the input image passes through the network once and then the objects can be detected. And we can get comprehensive learning. Except for the final layer, all other layers use ReLU as an activation relay.
The first 20 convolutional layers were tested by Imagenetto obtain an accuracy of up to 88%. The network is then trained for about 135 epochs on training and verification datasets of Pascal VOC.

5.5.2. Single Shot Detector (SSD):

SSD is a research paper published in the year [22] 2016 SSD speeds up the process of object discovery by eliminating the need for a network of area proposals.
To restore the decrease in resolution, the SSD implements some improvements including multi-scale features and .default boxes these improvements allow the SSD to match the resolution of the R-CNN Faster using lower resolution

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

images, which increases the speed even more. According to the following comparison, it achieves real-time processing speed and surpasses even the accuracy of the fastest R-Faster CNN.
An object can be detected using a two-part SSD:

- Extraction of landmark maps.
- Apply wrap filters to detect objects.

Each prediction consists of a boundary box and 21 scores for each category (one additional category for the absence of an object), and we choose the highest score as the category of the selected object. Conv4_3 makes a total of 4 × 38 × 38 predictions: four predictions per cell regardless of the depth of the landmark maps. As expected, many predictions do not contain any object. The SSD keeps category "0" to indicate that it does not contain objects.
Making multiple predictions with boundary boxes and confidence scores is called multibox.
We notice that we make much more predictions than the number of items found. So there are many more negative matches than positive ones. This creates an imbalance that harms training. We train the model to generalize the background space instead of detecting objects. However, the SSD still requires negative sampling in order to be able to figure out what constitutes a bad prediction. Therefore, instead of using all the negatives, we sort out those negatives by a calculated loss of confidence. The SSD selects the negatives with the highest loss and makes sure that the ratio between the negatives and the selected positives is at most 1:3. This leads to faster and more stable training.

6. Comparisons of object detection technologies

The following tables show a comparison of the previously reviewed algorithms
The review deals with the technique of zone suggestion, the possibility of inserting images of different measurements and the method of
The type of loss function used is also indicated, as is the case of the existence of a SoftMax layer and the language in which the algorithm is written.

Table 1, 2, 3: Comparison of some object detection technologies

Framework	**Proposal**	**Multi-scale Input**	**Learning Method**
R-CNN	Selective Search	No	SGD, BP
SPP-net	EdgeBoxes	Yes	SGD
Fast RCNN	Selective Search	Yes	SGD
Faster R-CNN	RPN	Yes	SGD
YOLO	-	NO	SGD
SSD	-	NO	SGD
YOLOv2	-	NO	SGD

Framework	Loss Function
R-CNN	Hinge loss (classification),Bounding box regression
Fast RCNN	Class Log loss + bounding box regression
Faster R-CNN	Class Log loss + bounding box regression
YOLO	Class sum-squared error loss + bounding box regression + object confidence + background confidence
SSD	Class softmax loss + bounding box regression
YOLOv2	Class sum-squared error loss + bounding box regression + object confidence + background confidence

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

Framework	Softmax Layer	End-to-end Train	Platform	Language
R-CNN	Yes	No	Caffe	Matlab
Fast RCNN	Yes	No	Caffe	Python
Faster R-CNN	Yes	Yes	Caffe	Python/ Matlab
YOLO	Yes	Yes	Darknet	C
SSD	No	Yes	Caffe	C++
YOLOv2	Yes	Yes	Darknet	C

7. Experience and results

The previously explained techniques will be tried on specific data sets to see the efficiency of each technique and its positions of superiority.

7.1. Training datasets:

The development of computer vision has led to an increased need for large-scale image datasets. A large amount of large explanatory data is the main reason behind the tremendous success of using deep learning technologies in object detection. The internet plays a vital role in building a comprehensive dataset to provide access to a wide range of images covering the breadth and diversity of objects. Five datasets are very popular in the field of general object discovery, namely HYPASCAL VOC2007, Pascal VOC 2012, ImageNet, Microsoft COCO and OpenImages. Table 3 summarizes the specifications and attributes of these data sets.

7.2. Results:

When trying the techniques that we have previously explained on the data, we arrive at the following results:

Table 4: Results of the Pascal VOC 2007 test on object detection techniques

Methods	Trained on	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN(Alex)	07	68.1	72.8	56.8	43	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	68.6	58.5
R-CNN(VGG16)	07	73.4	77	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1	66
Fast R-CNN	07+12	77	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82	76.6	69.9	31.8	70.1	74.8	80.4	70.4	70
Faster R-CNN	07	70	80.6	70.1	57.3	49.9	78.2	80.4	82	52.2	75.3	67.2	80.3	79.8	75	76.3	39.1	68.3	67.3	81.1	67.6	69.9
Faster R-CNN	07+12	76.5	79	70.9	65.5	52.1	83.1	84.7	86.4	52	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83	72.6	73.2
Faster R-CNN	07+12+COCO	84.3	82	77.7	68.9	65.7	88.1	88.4	88.9	63.6	86.3	70.8	85.9	87.6	80.1	82.3	53.6	80.4	75.8	86.6	78.9	78.8
SSD300	07+12+COCO	80.9	86.3	79	76.2	57.6	87.3	88.2	88.6	60.5	85.4	76.7	87.5	89.2	84.5	81.4	55	81.9	81.5	85.9	78.9	79.6
SSD512	07+12+COCO	86.6	88.3	82.4	76	66.3	88.6	88.9	89.1	65.1	88.4	73.6	86.5	88.9	85.3	84.6	59.1	85	80.4	87.4	81.2	81.6

Table 5: Summary of the results of the Pascal VOC 2007 test on object detection techniques

Methods	Trained on	mAP
R-CNN(Alex)	07	58.5
R-CNN(VGG16)	07	66
Fast R-CNN	07+12	70
Faster R-CNN	07	69.9
Faster R-CNN	07+12	73.2
Faster R-CNN	07+12+COCO	78.8
SSD300	07+12+COCO	79.6
SSD512	07+12+COCO	81.6

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

Table 6: Results of the Pascal VOC 2012 test on object detection techniques

Methods	Trained on	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN(Alex)	12	71.8	65.8	52	34.1	32.6	59.6	60	69.8	27.6	52	41.7	69.6	61.3	68.3	57.8	29.6	57.8	40.9	59.3	54.1	53.3
R-CNN(VGG16)	12	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7	82	74.8	76	65.2	35.6	65.4	54.2	67.4	60.3	62.4
Fast R-CNN	07++12	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73	55	87.5	80.5	80.8	72	35.1	68.3	65.7	80.4	64.2	68.4
Faster R-CNN	07++12	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5	70.4
YOLO	07++12	77	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8	57.9
YOLOv2	07++12+coco	88.8	87	77.8	64.9	51.8	85.2	79.3	93.1	64.4	81.4	70.2	91.3	88.1	87.2	81	57.7	78.1	71	88.5	76.8	78.2
SSD300	07++12+coco	91	86	78.1	65	55.4	84.9	84	93.4	62.1	83.6	67.3	91.3	88.9	88.6	85.6	54.7	83.8	77.3	88.3	76.5	79.3
SSD512	07++12+coco	91.4	88.6	82.6	71.4	63.1	87.4	88.1	93.9	66.9	86.6	66.3	92	91.7	90.8	88.5	60.9	87	75.4	90.2	80.4	82.2

Table 7: Summary of the results of the Pascal VOC 2012 test on object detection techniques

Methods	Trained on	mAP
R-CNN(Alex)	12	53.3
R-CNN(VGG16)	12	62.4
Fast R-CNN	07++12	68.4
Faster R-CNN	07++12	70.4
YOLO	07++12	57.9
YOLOv2	07++12+coco	78.2
SSD300	07++12+coco	79.3
SSD512	07++12+coco	82.2

8. Discussion of results

From the previous results we find the following:

- The negative of two-stage object detection technologies (zone suggestion) is the need for high computational power
- The structure of two-stage object detection technologies is more flexible and efficient than single-stage technologies.
- Single-stage object detection techniques (YOLO, SSD) require less time compared to two-stage techniques:
- Due to the CNN networks to extract less consuming features.
- Avoid preprocessing algorithms.
- There is no zone proposal stage.
- We note that the performance of single-stage object detection technologies is generally not excellent as these technologies find it difficult to detect small objects.
- When good gear is available to overcome the processing time, two-stage based technologies are the best option.

9. Conclusion

The success of object detection depends primarily on the technologies used. Single-stage technologies are clearer and faster, while two-stage ones are more accurate and efficient. The ultimate goal is to develop an object detection technology capable of accurately and efficiently identifying and recognizing thousands of new object classes in open real-world scenes. Thus, we need larger datasets with more categories, since the current standard datasets cover a few hundred categories of objects, which is much less than the categories recognized by man. In this paper, we reviewed the concept of object detection after providing a brief introduction to deep learning and convolutional neural networks. Then a download of the most important practical papers that launched a set of object detection algorithms based on deep learning was shown. We studied the structure of the most famous of these algorithms, the reason for their launch and the mechanism of their action. Then we tested this data and discussed the results to reach some criteria for using some object detection techniques.

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

References

[1]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[2]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
[3]. Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. Neurocomputing, 396, 39-64.

[4]. Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., & Tang, X. (2021). New generation deep learning for video object detection: A survey. IEEE Transactions on Neural Networks and Learning Systems.

[5]. Pitts, W., & McCulloch, W. S. (1947). How we know universals the perception of auditory and visual forms. The Bulletin of mathematical biophysics, 9(3), 127-147.

[6]. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.

[7]. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June).Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[8]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[9]. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

[10]. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[11]. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409 . 1556.

[12]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13]. O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[14]. Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification?. IEEE transactions on image processing, 24(12), 5017-5032.

[15]. Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[16]. Dalal, N., Triggs, B., & Schmid, C. (2006, May). Human detection using oriented histograms of flow and appearance. In European conference on computer vision (pp. 428-441). Springer, Berlin, Heidelberg.

[17]. Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[18]. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards realtime object detection with region proposal networks. Advances in neural information processing systems, 28.

[19]. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[20]. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117- 2125).

[21]. Pramanik, A., Pal, S. K., Maiti, J., & Mitra, P. (2021). Granulated RCNN and multi-class deep sort for multi-object detection and tracking. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(1), 171-181.

[22]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[23]. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[24]. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[25]. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024

detection. arXiv preprint arXiv:2004.10934.

[26]. Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

[27]. Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055.

[28]. Aziz, L., Salam, M. S. B. H., Sheikh, U. U., & Ayub, S. (2020). Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. IEEE Access, 8, 170461-170495.

DOI: https://doi.org/10.54216/NIF.030103
Received: April 29, 2023 Accepted: January 10, 2024