IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE STUDY OF YOLO AND SSD OBJECT DETECTION ALGORITHMS

[1]Peddinti Mounika, [2]Ch Lakshmi Narayana,

[3] Dr. Kondapalli Venkata Ramana

[1]M.Tech Student, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam.

[2]Research Scholar, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam.

[3]Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam.

Abstract: Object detection is one of the most important and challenging branches of computer vision, which has been widely applied in people's life, such as monitoring security, autonomous vehicle driving and so on, with the purpose of locating instances of semantic objects of a certain class. In order to understand the main development status of object detection pipeline, thoroughly and deeply, in this survey, we first examine the existing methods of typical detection models and describe the benchmark datasets. This project describes the role of deep learning techniques based on convolutional neural network for object detection. Deep learning techniques for state-of-the-art object detection systems are assessed in this project. A computer views all types of visual media as an array of numerical values. As a consequence, they require image processing algorithms to inspect contents of images. This project compares two major object detection algorithms: Single Shot Detection (SSD) and You Only Look Once (YOLO) to find the fastest and most efficient of the two. In this comparative analysis, using the COCO (Common Object in Context) dataset, the performance of these two algorithms is evaluated and their strengths and limitations are analysed based on parameters such as accuracy, precision and speed. From the analysis of results, it can be concluded that in an identical testing environment, YOLO-v3 performs better than SSD.

Keywords: ***YOLO, COCO, CNN, Deep learning.***

Introduction

Object detection has been attracting increasing amounts of attention in recent years due to its wide range of applications and recent technological breakthroughs. This task is under extensive investigation in both academia and real-world applications, such as security monitoring, autonomous vehicle driving, transportation surveillance, drone scene analysis, and robotic vision. Among many factors and efforts that lead to the fast evolution of object detection techniques, notable contributions should be attributed to the development of deep convolution neural networks and GPUs computing power. At present, deep learning model has been widely used in the whole field of computer vision, including general object detection and domain-specific object detection. Most of the state-of-the-art object detectors use deep learning networks as their backbone and detection network to extract features from input images (or videos), classification and localization respectively [1].

Object detection is a computer technology related to computer vision and image processing which deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include multi-categories detection, edge detection, and salient object detection, pose detection, scene text detection, face detection, and pedestrian detection etc. As an important part of scene understanding, object detection has been widely used in many fields of modern life, such as security field, military field, transportation field, medical field and life field. Furthermore, many benchmarks have played an important role in object detection field so far, such as Caltech, KITTI, ImageNet, PASCAL VOC, COCO, and Open Images V5[2,3]. Image classification and detection are important pillars of object detection. There is a plethora of datasets available. COCO is one such mostly used image classification domain. It is a benchmark dataset for object detection. It introduces a large-scale dataset that is available for image detection and classification [4]. This review article aims to make a comparative analysis of SSD and YOLO. The first algorithm for the comparison in the current work is SSD which adds layers of several features to the end network and facilitates ease of detection. While YOLO was developed by Joseph Redmon that offers end-to-end network [6]. In this article, by using COCO dataset as a common factor of the analysis and measuring the same metrics across all the implementations mentioned, the respective performances of the two above mentioned algorithms, which use different architectures, have been made comparable to each other. The results obtained by comparing the effectiveness of these algorithms on the same dataset can help gain an insight on the unique attributes of each algorithm, understand how they differ from one another and determine which method of object recognition is most effective for any given scenario [5].

In this work, a new direction of solving the bottleneck of proposals You Only Look Once (YOLO) will be discussed after the background descriptions of CNNs. YOLO v3 version is used in this project. By comparing YOLO with SSD on accuracy, speed, their advantages and shortages would be exposed. At last, performance of YOLO and SSD will be summarized [8].

1. Related Work

Object detection is an important topic of research in recent times. With powerful and effective learning tools available deeper features can be easily detected and studied. This work is an attempt to compile information on various object detection tools and algorithms so that a comparative analysis can be done and meaningful conclusions can be drawn to apply them in object detection. Literature survey serves the purpose of getting an insight regarding our work [1,2].

Wei Liu et al came up with a new approach of detecting objects in images using a single deep neural network. They named this method the Single Shot Multi Box Detector SSD. According to the team, SSD is a simple method and requires an object proposal as it is based on the complete elimination of the process that generates a proposal. It also eliminates the subsequent pixel and resampling stages. So, it combines everything into a single step. SSD is easy to train and is very straight forward when it comes to integrating it into the system. This makes detection of objects easier. The main feature of SSD is using multiscale convolutional bounding box outputs that are attached to several feature maps [8].

The paper by Pathak et al describes the role of deep learning technique by using CNN for object detection. The paper also accesses some deep learning techniques for object detection. The current paper states that deep CNNs work on the principle of weight sharing. It gives us information about some crucial points in CNN. In a recent research work by Chen et al, they have used anchor boxes for face detection and more exact regression loss function. They have proposed a face detector termed as YOLO face which is based on YOLOv3 that aims at resolving detection problems of varying face scales. The authors concluded that their algorithm out performed previous YOLO versions and its varieties [10]. The YOLOv3 was used in our work for comparison with other models.

In the research work by Fan et al, they have proposed an improved system for the detection of pedestrians based on SSD model of object detection. In this work the multi-layered system they introduced the Squeeze-and-Excitation model as an additional layer to the SSD model. The improved model has self-learning that further enhanced the accuracy of the system for small scale pedestrian detection. Experiments done on the INRIA dataset showed high accuracy [11]. This paper was used for the purpose of understanding the SSD model. Our comparison work was done using coco dataset.

2. Methodology

Data Collection:

This project includes You Only Look Once (YOLO) to find the fastest and most efficient in comparative analysis using the COCO (Common Object in Context) dataset. We used YOLO and SSD algorithms. This paper shows one of the best CNN representatives. You Only Look Once (YOLO), which breaks through the CNN family’s tradition and innovates a completely new way of solving the object detection with most simple and highly efficient way. SSD has fastest speed than YOLO. SSD achieved 59 fps and yolo with 45 fps. YOLO generated highest accuracy with 71% and SSD with 57%.

Data pre-processing:

Data preparation requires approximately 80% of time. Once data is gathered, it needs to be pre-processed, cleaned, constructed, and formatted in a style that comprehends and is able to work with. Deep learning tools should be used to analysed collected real-time data.

Training and Testing Data:

The proposed model needs to be trained and tested under various conditions by altering CNN and R-CNN parameters so that correctness can be obtained. In addition, we consider that the model’s accuracy is maximum. From the collected data will be used to train and test the model, respectively. In case of necessity, there must be provisions to improvise on the algorithm being used.

DSOD:

The proposed DSOD method is a multi-scale proposal-free detection framework similar to SSD. The network structure of DSOD can be divided into two parts: the backbone sub-network for extraction of feature and the front-end sub-network for prediction over multi-scale response maps. The backbone sub-network is a variant of the deeply supervised Dense Nets structure, which is composed of a stem block, four dense blocks, two transition layers and two transition w/o pooling layers. The front-end subnetwork (or named DSOD prediction layers) fuses multi-scale prediction responses with an elaborated dense structure.

3. Implementation

YOLOv3

YOLOv3(You Only Look Once, Version 3) is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. YOLO uses features learned by a deep convolutional neural network for object detection. Joseph Redmon and Ali Farhadi created Versions 1-3 of YOLO. Object classification systems are used by Artificial Intelligence (AI) programs to perceive specific objects in a class as subjects of interest. The systems sort objects in images into groups where objects with similar characteristics are placed together, while others are neglected unless programmed to do otherwise.

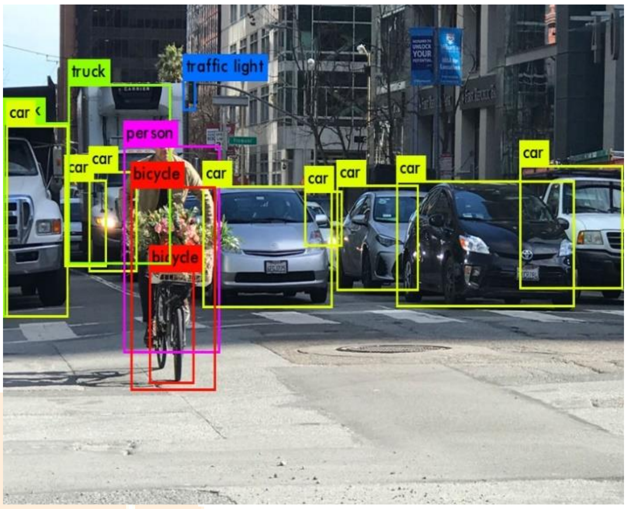

Figure 1. Object detection by YOLO algorithm

YOLOv3 Computer Vision Example

The first version of YOLO was created in 2016, and version 3, which is implemented in this paper, was made two years later in 2018. YOLOv3 is an improved version of YOLOv1 and YOLOv2. YOLO is implemented using the Keras or OpenCV deep learning libraries. The localization and classification heads were also united. Their single-stage architecture, named YOLO (You Only Look Once) results in a very fast inference time. The frame rate for 448x448 pixel images was 45 fps (0.022 s per image) on a Titan X GPU while achieving state-of-the-art mAP (mean average precision). Smaller and slightly less accurate versions of the network reached 150 fps. This new approach, together with other detectors built on light-weight Google's Mobile Net backbone, brought the vision (pun intended) of detection networks and other CV tasks on edge devices ever closer to reality.

CONVOLUTIONAL NEURAL NETWORK (CONVNET/CNN)

A Convolutional Neural Network (ConvNet/CNN) is a deep Learning algorithm. CNN can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlaps to cover the entire visual area. CNN is a particular type of feed-forward neural network in AI. It is used widely for image recognition. CNN represents the input data in the form of multidimensional arrays. It works well for a large number of labelled data. CNN extract the each and every portion of input image, which is known as receptive field. It assigns weights for each neuron based on the significant role of the receptive field. So that it can discriminate the importance of neurons from one another. CNN architecture consists of three types of layers: (1) convolution, (2) pooling and (3) fully connected [11].

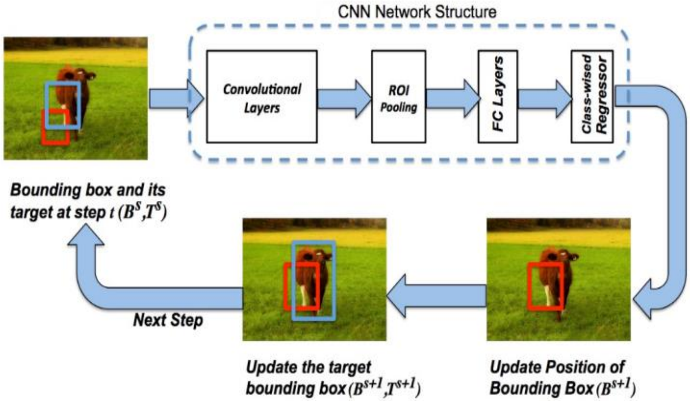
CNN Network Structure
Convolutional Layers
ROI Pooling
FC Layers
Class-wised Regressor
Bounding box and its target at step t (B^s,T^s)
Next Step
Update the target bounding box (B^{s+1},T^{s+1})
Update Position of Bounding Box (B^{s+1})

Figure 2. Architecture of CNN

YOLO

Shortly after that, You Only Look Once: Unified, Real-Time Object Detection (YOLO) paper was published by Joseph Redmon (with Girshick appearing as one of the co-authors). YOLO proposed a simple convolutional neural network approach which has both great results and high speed, allowing for the first-time real-time object detection [9].

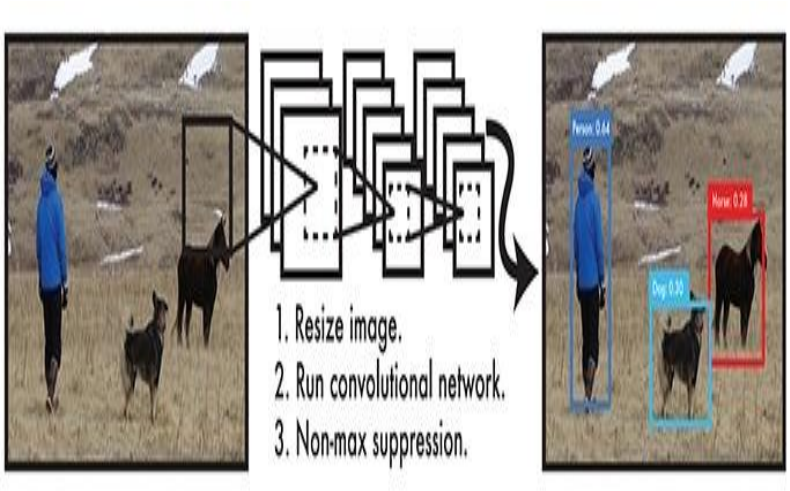

Figure 3. YOLO Architecture

SSD

When it comes to smaller objects, SSD’s performance is much worse as compared to YOLO. The main reason for this drawback, is that in SSD, higher resolution layers are responsible for detecting small objects. However, these layers are less useful for classification as they have lower-level features such as colour patches or edges, thereby reducing the overall performance of SSD. There is another limitation of this method which can be inferred from the complexity of SSD’s data augmentation, is that SSD requires a large amount of data for training purposes. It can be quite expensive and time consuming depending on the application[10].

4. Results

The application of deep learning algorithms is facilitated by pre-processing the data collected from multiple sources. Deep neural network-based object detection pipelines have four steps in general, image pre-processing, feature extraction, Classification and localization, post-processing. Firstly, raw images from the dataset can’t be fed into the network directly. Therefore, we need to resize them to any special sizes and make them clearer, such as enhancing brightness, color, contrast. Data augmentation is also available to meet some requirements, such as flipping, rotation, scaling, cropping, translation, adding Gaussian noise. Computer vision is an interdisciplinary field that has gained huge amounts of popularity in recent years. Another integral part of computer vision is object detection. Object detection aids in pose estimation, vehicle detection, surveillance etc. The difference between object detection algorithms and classification algorithms is that in detection algorithms, we try to draw a bounding box around the object of interest to locate it within the image. Also, you may not necessarily draw just one bounding box in the case of object detection, there could be many bounding boxes representing different objects of interest within the image and you would not know how many beforehand. Object detection is considered as foremost step in deployment of self-driving cars and robotics. In this paper, we demystified the role of deep learning techniques based on CNN for object detection. Deep learning frameworks and services available for object detection are also discussed in the paper. Benchmarked datasets for object localization and detection released in worldwide competitions are also covered. The pointers to the domains in which object detection is applicable has been discussed. State-of-the-art deep learning-based object detection techniques were evaluated and compared.

Object Detection in live video using SSD:

When user runs SSD algorithm for object detection, it automatically connects to webcam of laptop and detects objects in front of webcam for the trained objects in coco dataset. Comparing both SSD and YOLO in live video, though speed of SSD is greater than YOLO, yolo is more accurate than SSD.

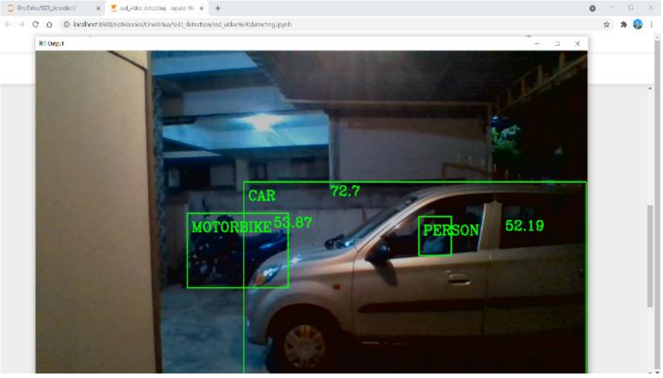

Figure 4. Object detection in live video using SSD

Object Detection in live video using YOLO:

When user runs yolo algorithm for object detection, it automatically connects to webcam of laptop and detects objects in front of webcam for the trained objects in coco dataset.

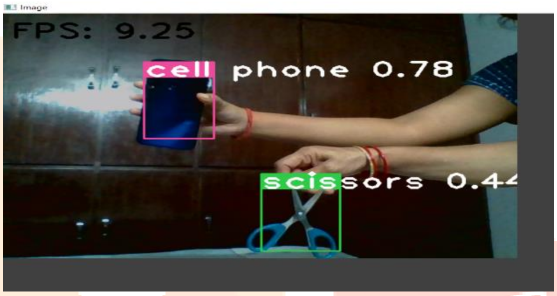

Figure 5. Object Detection in live video using YOLO

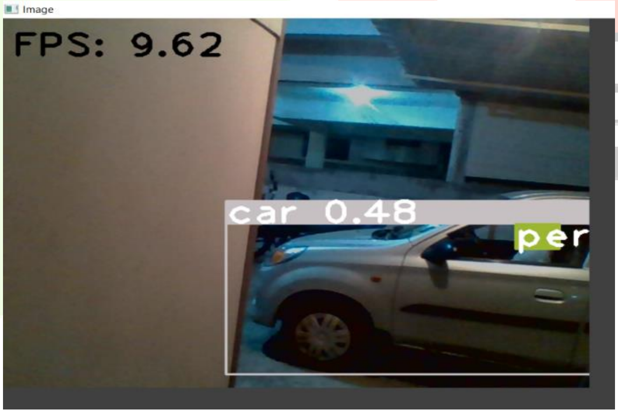

Figure 6. Object Detection in live video using YOLO

Object detection on image uploaded:

When any image is uploaded, the images that we have trained and tested in COCO dataset are detected. Detection of vehicle images like car, motorbike helps in vehicle detection.

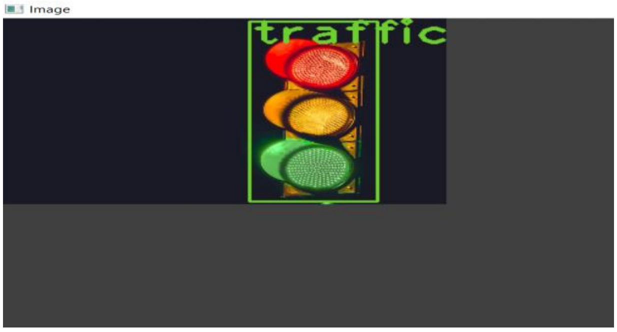

Figure 7: Object detection on image uploaded

Accuracy graph of YOLO and SSD:
From the below bar graph, we can say that yolov3 model performs better than SSD model.

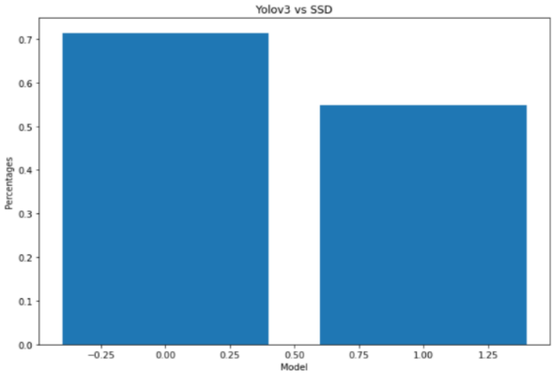

Figure 8: Accuracy graph of YOLO and SSD

YOLOv3 image classification vs object detection:

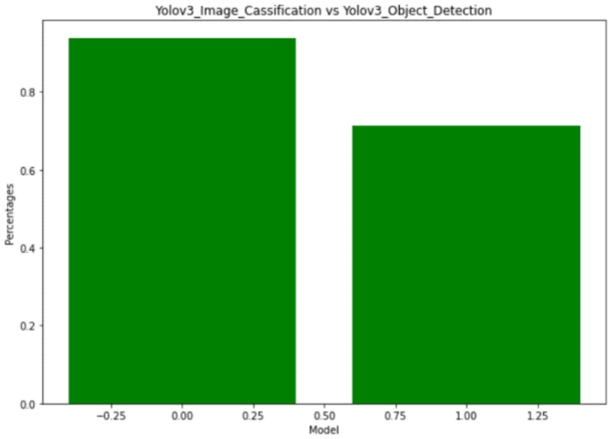

Figure 9. YOLOv3 image classification vs object detection

5. Conclusion

Object detection is considered as foremost step in deployment of self-driving cars and robotics. In this project, we demystified the role of deep learning techniques based on CNN for object detection. Deep learning frameworks and services available for object detection are also discussed in the paper. Benchmarked datasets for object localization and detection released in worldwide competitions are also covered. The pointers to the domains in which object detection is applicable has been discussed. State-of-the-art deep learning-based object detection techniques have been assessed and compared. This review article compared the latest and most advanced CNN-based object detection algorithms. Without object detection, it would be impossible to analyze the hundreds of thousands of images that are uploaded on internet every day. Technologies like self-driving vehicles that depend on real-time analysis are also impossible to realize without object detection. It was found that SSD is fastest than YOLOv3. Yolo-v3 is the one to pick if you need to analyze a live video feed. Meanwhile, SSD provides a good balance between speed and accuracy.

But YOLOv3 generates more accuracy than SSD. Hence, in conclusion, out of the two Object Detection Convolutional Neural Networks analyzed, Yolo-v3 shows the best overall performance. This result is similar to what some of the previous reports have obtained.

6. Future Work

The Future direction can be stated as follows. Due to infeasibility of humans to process large surveillance data, there is a need to bring data closer to the sensor where data are generated. This would result in real time object detection. Currently, object detection systems are small in size having 1-20 nodes of clusters having GPUs. These systems should be extended to cope with real time full motion video generating frames at 30 to 60 per second. Such object detection analytics need to be integrated with other tools using data fusion. The main problem is how to integrate processing into a centralized, powerful GPU for processing data obtained from various servers simultaneously and performs near real time detection analysis. To exploit the representational power of deep learning, large datasets over the size of 100 terabytes are essential. More than 100 million images are required to train the self-driving cars. Deep learning libraries should be augmented with prototyped environments in order to provide paramount throughput and productivity dealing with massive linear algebra-based operations. The datasets of image classification are widely available compared to that of object detection, the methods can be devised by which datasets meant for other tasks other than object detection would be applicable to be used for object detection. Existing methods are developed considering object detection as fundamental problem to be solved. There is scope to develop new design mechanisms capable of providing “Object Detection as a Service” in complex applications such as drone cameras, automated driving cars, robots navigating the areas such as planets, deep sea bases, and industrial plants where high level of precision in certain tasks is expected.

7. References

[1] S. Bell, C. Lawrence Zitnick, et al. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In CVPR, (2016).

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. In ICLR, (2015).

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, et al. Imagenet: A large-scale hierarchical image database. In CVPR, (2009).

[4] R. Girshick. Fast r-cnn. In ICCV, (2015).

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semanticsegmentation. In CVPR, (2014).

[6] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS,(2010).

[7] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In CVPR, (2016).

[8] B. Hariharan, P. Arbel ́aez, R. Girshick, and J. Malik. Hyper-columns for object segmentation and fine-grained localiza-tion. In CVPR, (2015).

[9] object detection YOLO algorithm

https://towardsdatascience.com/yolo-you-only-look-once-3dbdbb608ec4 (date accessed: august 26, 2021)

[10] Object detection SSD algorithm https://developers.arcgis.com/python/guide/how-ssd-works/

(date accessed: august 19, 2021)

[11] Deep neural networks for object detection https://machinelearningmastery.com/object-recognition-with-deep-learning/ (date published : February 2019).