JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

OBJECT DETECTION APPLICATION USING DEEP LEARNING

[1]Niraj Kumar Sahu, [2]Dewang Sharma, [3]Ishaan Shailesh Phaye

[1]Assistant. Professor, Dept. of Information Technology S.S.I.P.M.T, Raipur, India
[2]Student, Dept. of Information Technology S.S.I.P.M.T, Raipur, India
[3]Student, Dept. of Information Technology S.S.I.P.M.T, Raipur, India

Abstract: Object detection is a critical task in the field of Artificial Intelligence, Machine Learning, Deep Learning and computer vision with a wide range of applications such as tracking activity inside and outside a store or home, contact-less checkout, inventory management, video analytics among many other things. Convolutional neural networks (CNNs) are a strong tool that deep learning has developed as a way to use to detect objects with cutting-edge performance. In this study, we thoroughly analyze deep learning-based object detection. We address well-known CNN-based object identification algorithms including R-CNN, Fast R-CNN, and Faster R-CNN in addition to more contemporary methods like Single Shot MultiBox Detector (SSD), and You Only Look Once (YOLO). We also talk about the effects of object detection in physical-world applications, such as object recognition for automated driving and object detection for surveillance equipment. We conclude outline several potentials for deep learning-based object detection research, including interpretability, robustness to occlusion and environmental changes, and real-time object detection in resource-constrained situations.

Index Terms **- SSD, YOLO, Recognition, Detection, CNN.**

I. INTRODUCTION

A computer vision approach called object detection is used to find instances of semantic objects in digital photos or videos. Object detection is an artificial intelligence and deep learning related task of locating certain things inside a picture. In order to increase road safety, advanced driver assistance systems (ADAS) enable cars to recognize driving lanes or carry out pedestrian detection. For the challenge of item detection and recognition, much progress has been made in controlled environments. Nevertheless, the problem remains unresolved in unregulated environments, especially when objects are positioned haphazardly in a chaotic and obscured surrounding. For example, it may be uncomplicated to instruct a robot assistant to recognize the presence of a tea pot in an image when no other elements are present.

Take into account the level of complexity that would arise for a robot of this nature to pinpoint the machine on a kitchen surface that was filled with various items such as tools, devices, and utensils. This becomes extremely difficult in similar scenarios. To date, no viable solution has been discovered to address this issue. Throughout the past 20 years, a lot has been developed and researched in this field. Multidisciplinary approaches are frequently used in the study of object detection. Acquiring a comprehensive and up-to-date overview of the majority of state-of-the-art methodologies has become arduous and time-consuming due to the increasingly diverse and expansive nature of research innovations in this field.

Deep learning is a part of machine learning methods, which is based on artificial neural networks with representation learning. Deep learning algorithms are trained to identify and locate things in photos as part of object detection. This is accomplished by utilizing sizable datasets of annotated photos that include objects of interest and the bounding boxes that correspond to those objects. Typically, a convolutional neural network (CNN), which is intended to process and evaluate visual data, is the deep learning model for object detection. The annotated photos are used to train the CNN, which then learns to recognize the patterns and features that set one object apart from another. A set of bounding boxes enclosing the discovered objects and class labels are the model's output.

Object detection using deep learning involves two main stages: training and inference. During the training phase, the models are fed with large, annotated datasets, such as COCO (Common Objects in Context) or Pascal VOC, to learn the patterns and features of various objects. This process typically involves optimizing a loss function through backpropagation and gradient descent methods.

II. LITERATURE SURVEY

2.1 Early Approaches:

Deep learning has replaced earlier methods that depended on manually created features and classifiers for object detection. One of the first approaches for object detection utilizing Haar-like characteristics and an AdaBoost classifier was the Viola-Jones

algorithm, which was proposed in 2001. These methods, however, had poor accuracy and weren't resistant to changes in scale, orientation, and occlusion.

2.2 Rise of CNNs:

The breakthrough was the development of CNNs, which performed better than previous methods for classifying images. AlexNet, a deep CNN architecture presented by Krizhevsky et al. (Krizhevsky 2012), won the prestigious ImageNet Challenge (ILSVRC) by a wide margin, reigniting interest in deep learning for object detection. A number of CNN-based object identification techniques, including R-CNN, Fast R-CNN, and Faster R-CNN, which are widely regarded as ground-breaking innovations in the area, were developed as a result of inspiration from AlexNet.

2.3 R-CNN and its Variants:

Girshick et al's (Girshick 2014) proposal of region-based convolutional neural networks (R-CNN) presented the idea of region proposal methods. The region proposal generation and object classification phases of the object detection job are divided into two parts by R-CNN. It extracts prospective regions using selective search or other proposal generation techniques and then utilizes a CNN to categorize the objects contained within those regions. R-CNN attained cutting-edge performance at the time, but it was cumbersome because it had to independently review hundreds of regional proposal submissions.

The 2015 invention of Fast R-CNN enhanced R-CNN by sharing convolutional features across regions, resulting in quicker processing. Fast R-CNN used a Region of Interest (ROI) pooling layer to align the features of region proposals with fixed-size feature maps, which were then used for classification and bounding box regression. This method avoided passing cropped region proposals to CNN. The same year, a faster R-CNN was presented, which added a second Region Proposal Network (RPN) to provide region proposals and enabled end-to-end trainability of the object identification pipeline. This enhanced the object-detecting system's speed and precision even more.

2.4 Single-Shot Detectors:

To accomplish real-time object detection, single-shot detectors have been suggested as an alternative to region-based techniques. These techniques do away with time-consuming region proposal-generating processes by combining object categorization and region proposal generation into a single network. You Only Look Once (YOLO) and Single Shot Multi-Box Detector (SSD) are two well-known single-shot detectors.

In computer vision tasks, object detection often relies on the utilization of convolutional neural network architectures such as SSD MobileNet. Object detection is the process of identifying and localizing objects within an image or video. SSD MobileNet, a popular approach, employs deep learning techniques to perform object detection effectively. In SSD, the "SSD" MobileNet, which stands for Single Shot Detector, is a network that is particularly effective for real-time applications since it can detect objects in a given input image in a single pass. The "MobileNet" portion of the name alludes to the network's foundation in a lightweight architecture known as MobileNet, which is created to be quick and effective for deployment on mobile platforms or other platforms with limited resources.

.

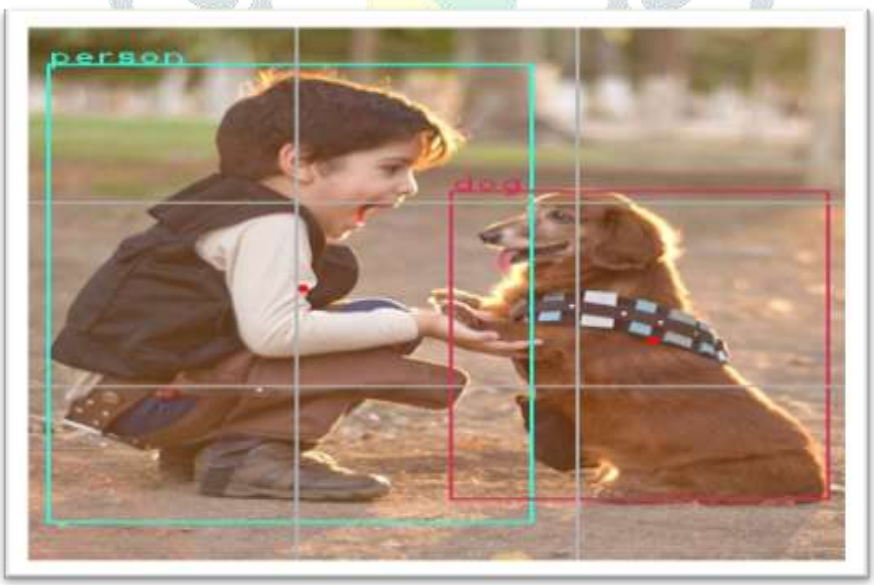

Fig.1 Picture to explain how object detection works

2.5 You Only Look Once (YOLO):

In YOLO, which was presented by Redmon et al. (Redmon 2015), the input image is divided into a grid, and object bounding boxes and class probabilities are predicted for each grid cell. Since it doesn't call for additional proposal creation processes, YOLO is renowned for its real-time processing capacity. Multiple feature maps with different resolutions can also be used to detect objects of various sizes. This method has been applied intensively in the field of deep learning-based models to identify objects of varying sizes. One such example is the SSD (Single Shot MultiBox Detector) method, introduced by Liu et al. (Liu 2016), which also utilizes a grid-based approach. SSD has been a popular choice for real-time object identification applications thanks to its competitive performance in terms of precision and processing speed.

III. Proposed System Architecture

3.1 Classification:

Classification refers to the process of identifying the category or class to which an object belongs. This is the most widely recognized issue with computer vision. Classification is used in object identification utilizing deep learning to identify the types of items present in an image, such as vehicles, people, or animals.

3.2 Localization:

Localization is the process of determining an object's exact location inside an image. When objects are found in an image, localization is used to identify their spatial coordinates, which are frequently shown as bounding boxes (e.g., x, y, width, and height).

3.3 Instance Segmentation:

The task of instance segmentation in computer vision is locating and separating distinct items within a picture. Instance segmentation, as opposed to object detection, which uses bounding boxes, gives pixel-level masks for each object, enabling more accurate object localization and separating overlapping objects.

3.4 Bounding Box Regression:

In object detection, bounding box regression is a technique used to improve the initial bounding box coordinates predicted by a model. It entails changing the bounding box's coordinates to more closely match the actual location of the object, usually through the use of regression techniques.

3.5 Intersection over Union (IoU):

The Intersection over Union (IoU) statistic is used to evaluate the accuracy of object detection systems. It computes the amount of overlap between an object's expected and actual ground truth bounding boxes. IoU is frequently used to evaluate the precision of object localization and is calculated as the area of the overlap(intersection) of the two bounding boxes parted by the union of their area.

$$Intersection\ over\ Union(IoU) = \left[\frac{Area\ of\ Overlap}{Area\ of\ Union}\right] \quad (1)$$

In conclusion, object detection using deep learning entails the tasks of classification (determining the type of objects), localization (estimating the location of objects with bounding boxes), instance segmentation (delineating specific objects with pixel-level masks), bounding box regression (refining bounding box coordinates), and evaluating the accuracy of predictions using metrics like Intersection over Union (IoU).

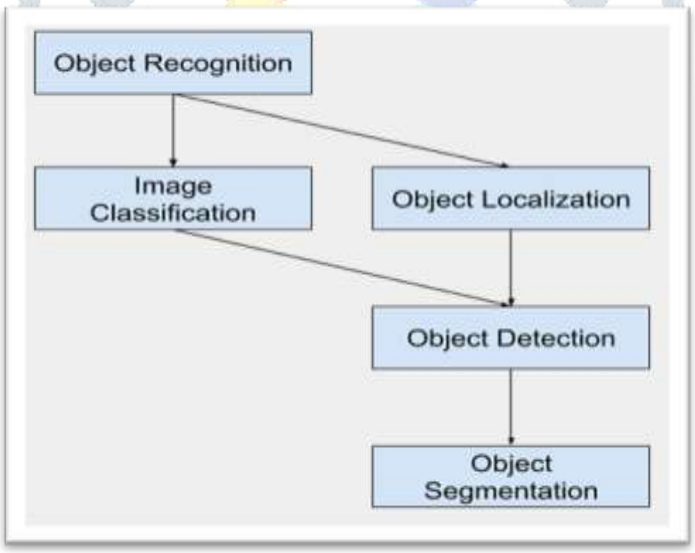

Fig.2 Overview of Object Detection

IV. Implementation

4.1 Classification:

We attempted to recognize items using the conventional method of the application of OpenCV libraries and a pre-trained deep learning model. Our pre-trained model was created using the SSD mobilenet method. We employed pre-trained models from Mobilenets to put the SSD strategy into practice. This method may categorize labels based on the learned model. We used the MS-COCO dataset, which had 91 classifications, as shown in fig. The static images and the input video are loaded as input, and each frame is scaled to a predetermined size of 300x300 pixels before being converted to a single frame input drop. Two files—one for setup and the other for weights—are included with our pre-trained models. Consequently, the model illustrates the division of neurons into two categories within a neural network: configuration and weights. We used the 2017 edition 2969 which contains a brand-new, unlabelled dataset of 123K images.

```
'Person', 'Bicycle', 'Car', 'Motorcycle', 'Airplane', 'Bus', 'Train',
'Truck', 'Boat', 'Traffic Light', 'Fire Hydrant', 'Street Sign', 'Stop Sign',
 'Parking Meter', 'Bench', 'Bird', 'Cat', 'Dog', 'Horse', 'Sheep', 'Cow',
'Elephant', 'Bear', 'Zebra', 'Giraffe', 'Hat', 'Backpack', 'Umbrella',
'Shoe', 'Eye Glasses', 'Handbag', 'Tie', 'Suitcase', 'Frisbee', 'Skis',
'Snowboard', 'Sports Ball', 'Kite', 'Baseball Bat', 'Baseball Glove',
'Skateboard', 'Surfboard', 'Tennis Racket', 'Bottle', 'Plate', 'Wine Glass',
'Cup', 'Fork', 'Knife', 'Spoon', 'Bowl', 'Banana', 'Apple', 'Sandwich',
'Orange', 'Broccoli', 'Carrot', 'Hot Dog', 'Pizza', 'Donut', 'Cake', 'Chair',
 'Couch', 'Potted Plant', 'Bed', 'Mirror', 'Dining Table', 'Window', 'Desk',
'Toilet', 'Door', 'Tv', 'Laptop', 'Mouse', 'Remote', 'Keyboard', 'Cell
Phone', 'Microwave', 'Oven', 'Toaster', 'Sink', 'Refrigerator', 'Blender',
'Book', 'Clock', 'Vase', 'Scissors', 'Teddy Bear', 'Hair Drier',
'Toothbrush', 'Hair Brush']
```

Fig.3 MS-COCO dataset

4.2 SSD-Mobilenet Architecture:

The MobileNet network serves as the SSD (Single Shot MultiBox Detector) detector's most crucial and dependent network model. The SSD methodology and speed rating precision are improved in real-time using the MobileNet method. With this method, multiple objects can be detected in a single shot. The SSD utilizes a neural network architecture for object detection. Before combining the default box to hold the contents, this network swiftly checks it for the presence of numerous object classes. This also enables simultaneous categorization and localization tasks to be performed concurrently. Non-Maximum Suppression (NMS) was also employed to eliminate most of these bounding boxes. Additionally, this network supports a range of models with different natural bond sizes and resolutions. This was done either because they contained an object that was already detected by another bounding box with a high confidence level or because their own confidence score was low.

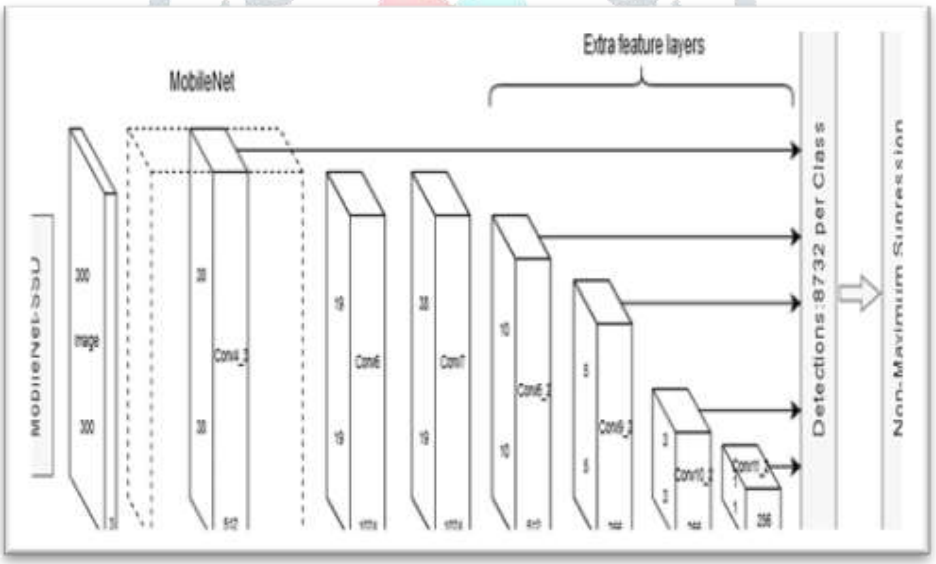

Fig.4 Overview working of SSD-Mobilenet

4.3 Neural Network API (NNAPI):

A neural network API (Application Programming Interface) in deep learning is a set of tools, protocols, and routines that allow developers to easily create and train artificial neural networks. Neural network APIs provide a high-level abstraction layer that simplifies the process of building and configuring complex neural networks by providing pre-defined network architectures, optimization algorithms, and other tools. Neural network APIs are typically part of a larger deep learning framework, which provides additional functionality such as data preprocessing, model evaluation, and deployment. Examples of popular deep learning frameworks that include neural network APIs are TensorFlow, PyTorch, and Keras.

V. Result Analysis

The performance of object detection using with SSD MobileNet and the MS COCO dataset can be analyzed using various metrics such as mean average precision (mAP), precision, recall, and F1 score. When one the key metrics used to evaluate the performance and accuracy of object detection algorithms is mAP (mean Average Precision), which is used for both precision and recall. If the mAP score is higher, the better the model's ability to detect and classify objects accurately. The performance and accuracy of an

object detection application can be more enhanced by using techniques like data augmentation, transfer learning, and model optimization. traditional methods by a significant margin.

5.1 Mean Average Precision (mAP):

A popular statistic for assessing object detection models is mAP. The average precision across all memory levels is measured. By generating the precision-recall curve and averaging the area under the curve (AUC) values for each class, you may get mAP. Better performance is indicated by a higher mAP value.

$$mAP = \left[\frac{1}{N}\sum_{k=1}^{k=n} AP_k\right] \qquad (2)$$

AP_k = The AP of class k

n = the number of classes

5.2 Advanced Metrics:

These are of 3 types: F1 score, Precision, and Recall. The mentioned metrics can be employed to assess the model's performance on a per-class basis. In all positive forecasts, precision is the percentage of genuine positives. Recall is the ratio of correctly classified positive samples to the total number of positive samples. The F1 score is the harmonic mean of recall and precision. Better performance is indicated by a higher F1 score.

$$F_1 = 2 \times \left[\frac{precision \times recall}{precision + recall}\right] \qquad (3)$$

$$Precision = \left[\frac{TP}{TP + FP}\right] \qquad (4)$$

$$Recall = \left[\frac{TP}{TP + FN}\right] \qquad (5)$$

In our research, we used the MS-COCO dataset to verify the proposed pre-trained Mobilenet- SSDv3 detector's detection results. A 512x512 RGB color filter is the input image format for our network model. However, the processing speed and detection precision of the proposed detector can be significantly increased. Our object detection technique has a maximum detection rate of 30 frames per second. During our testing, the SSD approach successfully displayed both indoor and exterior stream video frames using a camera and static graphics, although the positioning of the items varied between two consecutive frames.

During our testing, the objects' locations varied between two frames. The algorithm and the video captured by the webcam reduce the size of each frame to 300 × 300 pixels. By employing more default boxes, as they can have a greater impact, and unique boxes for each place; for various classes and confidence levels and for multiple objects at the same time, the SSD can generate a variety of bounding boxes. This recommended single-shot multi-box detection method makes use of frame difference.

Fig.5 Object Detection on Multiple Objects

VI. Conclusion

In recent past years, deep learning has garnered significant attention in the field of object recognition. In this study, our aim was to identify objects captured by a camera. To achieve this, we developed a model that leverages the pre-training of MobileNet and Single Shot Multi-Box Detector. Given the challenges associated with reading frames from a camera, we focused on achieving a high frames-per-second solution to mitigate input and output issues. Our proposed approach involves a user-friendly network design that enhances feature extraction using the Mobilenet-v3 backbone network. By combining Mobilenet and SSD models, we expand the feature map of the input and enhance the feature map of the input image, thereby improving the detection accuracy and performance of the back-end detection network. Through precise object localization in the x and y-axis coordinates of the frame, we are able to detect and identify objects more accurately and sequentially based on the results of our testing.

The study further encompasses experimental results involving various strategies for item detection and identification, along with a comparative analysis of their effectiveness. This presents a significant advantage for x86 hardware with limited resources. Based on the conducted experiments, the suggested Mobilenet-SSDv3 detector demonstrates a noteworthy enhancement in detection performance while retaining the original MobileNet-SSD detector's advantage of fast processing. This is done by combining two methods: efficient, threaded video streaming with OpenCV and deep learning with OpenCV for object detection. Thus an object detection method with high accuracy has been developed utilizing the SSD-MobileNet model, enabling its effectiveness across various cameras. In its dataset, our system can identify objects like cars, motorcycles, bottles, couches, and more. In order to help the community and make the system more interesting and appealing, this study aims to develop an autonomous system.

.

References

[1] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In Computer Vision– ECCV 2008, Springer, 2008.

[2] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2888-2897. Bounding Box Regression With Uncertainty for Accurate Object Detection.

[3] Object Detection Combining Recognition and Segmentation [Fudan University, Shanghai, Wang1, Jianbo Shi2, Gang Song2, and I-fan Shen] 2018.

[4] Shraddha, M., Supriya, M. Moving object detection and tracking using convolutional neural networks IEEE Xplore ISBN:978-1-5386-2842-3.

[5] Wu, J., Leng, C., Wang, Y., Hu, Q., J. Cheng, J. Quantized convolutional neural networks for mobile devices. arXiv preprint arXiv:1512.06473, 2015.

[6] Dhillon, Anamika, and Gyanendra K. Verma. "Convolutional neural network: a review of models, methodologies and applications to object detection." Progress in Artificial Intelligence 9.2 (2020).
[7] Object Detection With Deep Learning: A Review January 2019 IEEE Transactions on Neural Networks and Learning Systems PP(99):1-21 Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu Impaired U.S. Patent No. 9,488,833.8 Nov. 2016.
[8] Christian Szegedy, Alexander Toshev, and Dumitru Erhan, "Deep Neural Networks for Object Detection," IEEE, 2007.
[9] Hannes Schulz and Sven Behnke. Object-class segmentation using deep convolutional neural networks. In Proceedings of the DAGM Workshop on New Challenges in Neural Computation, 2011.
[10] Brent A. Griffin, Jason J. Corso; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1397-1406. Depth From Camera Motion and Object Detection.
[11] Kanimozhi, S., Gayathri, G., & Mala, T. (2019, February). Multiple Real-time object identification using Single shot Multi-Box detection. In the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS) IEEE.
[12] Adami Fatima Zohra, SalmiKamilia, Abbas Faycal, and SaadiSouad, "Detection And Classification Of Vehicles Using Deep Learning," International Journal of Computer Science Trends and Technology (IJCST), vol. 6, 2018.
[13] Chiu, Y. C., Tsai, C. Y., Ruan, M. D., Shen, G. Y., & Lee, T. T. (2020, August). Mobilenet-SSDv2: An improved object detection model for embedded systems. In 2020 International Conference on System Science and Engineering (ICSSE) IEEE.
[14] Rafael Padilla; Sergio L. Netto; Eduardo A. B. da Silva. A Survey on Performance Metrics for Object-Detection Algorithms (2011).
[15] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, 2014.
[16] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.