Performance Analysis of Prominent Object Detection Algorithms

by

Mst.Sumiya Siddika (Exam Roll: 192292)
Md. Masud Rana (Exam Roll: 192332)
Shariful Islam (Exam Roll: 192337)

A Researech Project Report submitted to the
Institute of Information Technology
in partial fulfillment of the requirements for the degree of
Bachelor of Science in
Information and Communication Technology

Supervisor: Dr. Fahima Tabassum
Professor

Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342

DECLARATION

We at this moment declare that this thesis is based on the results found by ourselves. Materials of work found by other researchers are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.

Signature of the Candidate
Mst.Sumiya Siddika

Signature of the Candidate
Md. Masud Rana

Signature of the Candidate
Shariful Islam

CERTIFICATE

This is to certify that the thesis entitled Performance Analysis of Prominent Object Detection Algorithms has been prepared and submitted by Mst. Sumiya Siddika, Md. Masud Rana, Shariful Islam in partial fulfilment of the requirement for the degree of Bachelor of Science (hon's) in Information and Communication Technology on November 15, 2023.

Dr. Fahima Tabassum
Supervisor

Accepted and approved in partial fulfilment of the requirement for the degree Bachelor of Science (honors) in Information and Communication Technology.

Dr. Fahima Tabassum
Chairman

Dr. Mohammad Shahidul Islam
Member

Dr. M. Mesbahuddin Sarker
Member

Dr. Hasanul Kabir
External

ACKNOWLEDGEMENTS

We feel pleased to have the opportunity to express our heartfelt thanks and gratitude to those who all rendered their cooperation in making this report. This thesis is performed under the supervision of Dr. Fahima Tabassum, professor of the Institute of Information Technology (IIT), Jahangirnagar University, Savar, Dhaka. During the work, she has supplied us with a number of books, journals, and materials related to the present investigation. Without her help, kind support and generous time span she has given, we could not have performed the project work successfully in due time. First and foremost, we wish to acknowledge our profound and sincere gratitude to her for her guidance, valuable suggestions, encouragement, and cordial cooperation. We express our utmost gratitude to Dr. Fahima Tabassum, Professor, IIT, Jahangirnagar University, Savar, Dhaka, for her valuable advice that has encouraged us to complete the work within the time frame. Moreover, we would also like to thank the other faculty members of IIT who have helped us directly or indirectly by providing their valuable support in completing this work. We express our gratitude to all other sources from where we have found help. We are indebted to those who have helped us directly or indirectly in completing this work. Last but not least, we would like to thank all the staff of IIT, Jahangirnagar University, and our friends who have helped us by giving their encouragement and cooperation throughout the work.

ABSTRACT

In the modern world, computer vision technology has become an important field of study for Internet applications. Currently, object detection forms the basis of many visual tasks and has become a fundamental problem in computer vision. Whether we need to identify the categories or understand the interaction between text and graphics, it provides reliable information. This article compares several object detection algorithms. Using underlying deep models, researchers have experimented extensively and contributed to the performance increase of identifying objects and related tasks including object classification, localization, and segmentation throughout the past ten years due to the rapid evolution of deep learning. There are two types of object detectors: single-stage and two-stage object detectors. Single-stage detectors, on the other hand, focus on all geographic areas suggested for possible detection of objects via comparatively simpler architecture in a single shot, while two-stage detectors primarily focus on methods to identify selective region proposals through complex architecture. Any object detector's performance is measured by its deduction time and accuracy of detection. Comparatively two-stage object detectors perform better in terms of detection accuracy than single-stage detectors. The four most widely used deep-learning algorithms will be stated in this article. We'll contrast the accuracy and efficiency of those algorithms. The algorithms are D2Det (The Deformable two-stage detector), YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector), and CNN (Convolutional Neural Network). CNNs are a particular kind of neural network used to process data that looks like a grid, like pictures and videos. To automatically recognize and remove features from input data, it makes use of convolutional layers. The benefits of flexible convolutional networks and two-stage detectors are combined in the object detection algorithm which is known as D2Det. Using a dense prediction network, it first produces region suggestions. With just one forward pass, the real-time object detection algorithm YOLO can accurately predict boundaries and class probabilities from an input image. The input image is divided into a grid by YOLO, which then predicts bounding boxes and class probabilities for each grid cell. A different real-time object detection algorithm, called SSD, combines several boundaries, ratios, and scales at various neural network levels. It does this without asking for region proposals by predicting bounding box coordinates and object classes. It offers recommendations for the future development pattern and research of object detection by analyzing the state of object detection network research at the moment.

Keywords: CNN, YOLO, D2det, SSD.

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
YOLO	You Only Look Once
D2Det	The Deformable two-stage detector
SSD	Single Shot Multibox Detector

LIST OF FIGURES

Figure

3.1 ReLU function image 11
3.2 Architecture of CNN 12
3.3 Flow chart of CNN network 13
3.4 Proposed system model 20

TABLE OF CONTENTS

DECLARATION ii

CERTIFICATE iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

CHAPTER

I. Introduction 1

1.1 Background 1
1.2 Overview 3
1.2.1 Impact of the study 3
1.3 Problem Statement 4
1.4 Objective 4
1.4.1 Scope of the study 4

II. Literature Review 5

2.1 Related Work 5

III. Methodology 10

3.1 Basic theory and Algorithms analysis 10
3.2 Proposed System Model 20

IV. Conclusion and Future Work 22

4.1 Conclusion 22

4.2 Future work 22

References 23

CHAPTER I

Introduction

1.1 Background

A computer vision task called object detection includes finding and recognizing things within a picture. Object detection aims to build bounding boxes around objects in order to offer information about their exact positions in addition to identifying which objects are there in the visual input. We need object detection in visual understanding, automation, Security and Surveillance, Medical Imaging, Quality Control and Manufacturing. Through object detection, machines can identify and locate items in photos. When it comes to automating procedures that call for interaction with the real world, object detection is essential. In order to locate and identify anatomical structures, growths, and anomalies in X-rays, MRI scans, and other medical pictures, object identification is used in medical imaging. Object detection is very important in those sectors. There are different techniques that are used in object detection. Some popular methods are given below:

CNN (Convolutional Neural Network): A subclass of deep neural networks called CNNs is specially made to handle organized grid data. They have been widely used in the processing, analysis, and recognition of images and videos. CNNs use a type of perceptron with multiple layers that are intended to require very little preprocessing. They immediately pick up the structure and can learn directly from raw data. CNNs have made a major contribution to computing's advancement.

YOLO (You Only Look Once): YOLO, a real-time object detection system, is among the most widely used and fastest algorithms for this kind of task. YOLO organizes photos into a grid and predicts bounding boxes and probability classes for each grid cell. This leads to the prediction of bounding boxes and class probabilities for the entire image using a single network evaluation. There are multiple versions of YOLO, including YOLOv1, YOLOv2, YOLOv3, and YOLOv4, each with different speed and accuracy enhancements.

D2Det (The deformable two-stage detector): This algorithm combines the strengths of deformable convolutional layers with the two-stage object detection approach to achieve improved accuracy in detecting and localizing objects, especially in cases where objects exhibit deformations, occlusions, or variations in scale.

SSD (Single Shot Multibox Detector): Another well-liked object detection method with a track record for single-shot detection is SSD. In a picture, it simultaneously predicts multiple bounding boxes and their class probabilities at different scales. SSD effectively recognizes objects that have different scales and aspect ratios by utilizing a variety of default bounding boxes with different ratios. SSD is known for achieving a balance between precision and speed, which qualifies it for real-time applications.

Anchor Boxes: Default bounding boxes, sometimes referred to as anchor boxes, are used to predict object sizes and locations at various scales and aspect ratios. They give a hint about what types of objects to look for in different parts of the picture.

Backbone Networks: The architecture of a backbone network has an important impact on how well object detectors work. Typical options include ResNet, VGG, Inception, and EfficientNet architectures.

In comparison with other methods, SSD is the quickest object-detecting system with the lowest accuracy, whereas YOLOv3 is the most accurate but slowest. While YOLOv2 is faster than YOLOv3, it is less accurate. YOLOv3 is the greatest option for object detection in captured photos and videos since it has the highest accuracy in object detection. Every one of these systems loses speed and accuracy, therefore the best solution for a given application is determined by the needs of the application.

1.2 Overview

In this report, we will discuss about object detection, the techniques that we will use and the problems of our proposed topic. We will discuss different algorithms that are related to our topic and their parameters. Here we will also find the impacts of our project. We will discuss about the algorithms that will be used in this thesis and their execution.We will give a proposed model which we will maintain in case of doing this thesis. We will give what we will do in future.

1.2.1 Impact of the study

1.2.1.1 National Impact :

Creating techniques to identify several things in a single picture or video frame. Increasing object detection is essential for both public safety and national security. In security systems, airports, and public areas, it is used to identify and track items, people, and possible threats. Object detection in agriculture helps in automating farming operations, detecting crop diseases, and optimizing crop management. Food security and agricultural output may rise as a result. Object detection is a tool used by investigators and police departments for traffic monitoring, missing person searches, and criminal investigations. It helps keep law and order and solve crimes.

1.2.1.2 International Impact:

In international security and defense, object detection technologies are employed to guard airports, borders, and other critical infrastructure. It helps identify and stop terrorist threats as well as other security-related issues. Object detection technologies monitor crop conditions and optimize agricultural operations to help global food production and security objectives.

1.3 Problem Statement

The main purpose of our study is to detect objects using four popular algorithms named CNN, YOLO, SSD and D2det. Here we will also calculate performance, accuracy so that we can state a comparison among those algorithms.

1.4 Objective

The main objectives are to:

- Use four algorithms named CNN, YOLO, SSD and D2Det in object detection.
- Discuss those algorithms in a detailed way.
- Evaluate and compare selected algorithms on the basis of their performance, speed, and accuracy.

1.4.1 Scope of the study

- Real-time Object Detection: Investigating methods for real-time or almost real-time object identification for robots and driverless vehicle applications.
- Small Object Detection: Tackling the difficulties associated with finding small objects in pictures, which are frequent in fields including medical imaging, satellite imaging, and microscopy.
- Multi-Object Detection: Developing techniques to identify several things in a single picture or video frame. Expanding item detection into highly populated areas.

CHAPTER II

Literature Review

2.1 Related Work

1. In the first paper, the author gave the information that 2 types of object detection methods are found. One-stage methods like YOLO, SSD and two-stage methods for instance R-CNN, fast R-CNN, and faster R-CNN. CNN is more effective than the traditional handmade image extraction method from images. But CNN is more time-consuming than the YOLO method. The limitation of the YOLO algorithm is it can't detect overlapping small objects.SSD means single shot multibox detector has a significant difference from the YOLO algorithm. SSD uses direct detection of convolution and can overcome problems of the YOLO algorithm [10].

2. The YOLOv5 model produces results with an exceptionally good visualization function. This study shows that the TSR in the YOLOv5 experiment is remarkably accurate. The definitions of "road bump," "crosswalk," "give way," and "no entry" are given in detail in this document. The "No U-turn" method produced the lowest precision of 0.94. Almost eight classes have values that are all over 90.00%, demonstrating YOLOv5's exceptional TSR performance in our dataset.[19].

3. In this paper, Objects that may be a risk to the safety of an autonomous vehicle when driving on a road are classified as hurdles, cars, and passengers. The advantage of YOLOv4, a typical one-stage detector technique, is its quick detection speed[2].

4. In this paper for object detection experimentation, there are three types of photos utilized process: 409 for training, 46 for validation, and 51 for testing. Important metrics that show how accurately object detection algorithms recognize objects

are AP and mAP. The accuracy of the YOLOv4 model is 93.97%, while YOLO-GD obtains 97.38%, with the larger value of AP or mAP[17].

5. This is mainly a review paper. The three-stage object detectors—RCNN, Fast-RCNN, and Faster-RCNN—as well as their significant applications were studied in this paper. This research reviewed in detail single-stage object detectors, in particular YOLO objects, their architectural developments, and their loss function[3].

6. This is a review paper that compares the performance of various object detectors on PASCAL VOC 2012 and Microsoft COCO datasets. Those models are compared on average precision (AP) and processed frames per second (FPS) at inference time. This paper intentionally compares the performances of detectors on similarly sized input images, where possible, to provide a reasonable account[18].

7. This study describes the YOLOv5, a deep learning-based bug detector. For the purpose of training, validation, and testing, this model created a new twenty-three classes IP-23 dataset. The model that produced the most success, YOLOv5x, was determined to be notable. The YOLOv5x model, which was designed for this project and trained with specific parameters, produced an average precision value of 98.3%, recall value of 97.8%, precision value of 94.5%, and F1 score of 96% in terms of detection rate[1].

8. An improved YOLO technique for target detection in high-resolution zoom-sensing photos is presented in this study. It uses the SLIC superpixel segmentation technique to distinguish between light and dark areas, hence addressing issues like image blur and distortion. In order to deal with the unique properties of the image, the suggested technique modifies the vertical grid number of the YOLO network structure. Experiments done with multiple datasets show that performs regular YOLO and other popular algorithms in terms of accuracy and real-time performance[11].

9. This paper describes the use of machine learning and DWT for human face recognition. This paper employs four distinct algorithms: the principal component analysis (PCA) error vector, the PCA eigenvector, the CNN eigenvector, and the Linear Discriminant Analysis (LDA) eigenvector. The four results are then combined using the fuzzy system and the entropy of detection probability. The paper's combined approach produces a recognition rate of 93.34% in the best scenario and 89.56% in the worst[12].

10. In this paper author showed that Normal machine learning in object detection does not perform well but combination of YOLO and SSD can accurately detect objects and CNN-based YOLO enhances processing time. Also, we find that YOLOV5 is faster than YOLOv3 that's why YOLOv5 is used over YOLOv3[5].

11. In the paper named Hyperspectral Anomaly Detection Using Deep Learning, the author helped us by providing important information about anomaly detection from images using deep learning. CNN one of the deep learning methods has better fault tolerance, adaptability, and strong self-learning ability. CNN can extract features from images automatically. End-to-end cube CNN is better than pixel-based CNN but it can be affected by noise and inference[13].

12. Multiscale object recognition in Synthetic Aperture Radar (SAR) pictures is growing as an important area of research in SAR image interpretation. In this paper, the author proposed an approach named SARFNet. SARNet has the highest detection accuracy over other state-of-the-art methods. This paper also gave information adding saliency information in SSD algorithm guides SSD to understand the Salient feature of the SAR target [14].

13. This paper is about object Instance Segmentation. To overcome the problems of object instance segmentation detection-based Methods and single-stage methods are used. The detection-based method has the highest accuracy and single-stage methods have faster speed. The future work of this paper is to segment object instances accurately in bad weather conditions like rain, snow, etc and multiscale object segmentation [6].

14. The latest developments in computer vision are covered in this excerpt, with a focus on deep learning methods. It highlights the importance of tasks like object detection, object recognition, and image classification while highlighting the critical role that convolutional neural networks (CNNs) play in these processes. The evolution of object recognition from single to multi-object recognition is highlighted in the text, along with the use of deep learning in this field. It also discusses how deep learning-based methods—like RCNN—help achieve greater accuracy across a range of datasets[4].

15. In the passage, the significance of object detection in computer vision is discussed, and the "DeepMultiBox" detector is presented as a possible fix for the computing difficulties associated with the exhaustive search method. This detector uses a single Deep Neural Network (DNN) in a class-agnostic manner to produce

a finite number of bounding boxes as object possibilities. It highlights the novel use of regression to define object detection, which enables the net to generate a confidence score for every projected box. Compared to conventional techniques that score features inside specified boxes, this unique approach is different. The paper's unique loss function, which makes it easier to train bounding box predictors inside the network, is one of its main contributions. Through the resolution of an assignment issue involving ground truth boxes and predictions, as well as the updating of matched box locations, confidences, and matched box coordinates, confidences, and underlying features, the model is tailored toward precise localization [4].

16. Active research has been done on object detection, and virtually every few months, fresh, cutting-edge findings are published. There are still a lot of unresolved issues, though. Below, we go over a number of unresolved issues and potential paths. [16].

17. This paper is about road object detection using deep learning. Using these algorithms it may detect objects in most of the cases [7].

18. Here, they suggested a hybrid system for object recognition and pedestrian detection that combines both CNN and SVM. The appearance of items in a real environment fluctuates because of factors including shadows, partial occlusion, light variations, and background clutter. This could result in incorrect object detection and recognition in autonomous car applications, which could have disastrous consequences. In this study, they proposed an LM-CNN-SVM system to address these issues. They employed a new CNN architecture with nine layers in addition to a pre-trained AlexNet architecture in our system. In order to extract the discriminative features of each patch, they partitioned the entire image into patches using CNN. Next, in order to de-correlate and decrease the features that they had retrieved from the CNN, we used PCA.

Finally,they imported them into the SVM classifiers' input to improve the system's capacity for generalisation. Using a majority voting procedure, they then successfully fused the images.. [15].

19. In this paper, they proposed a novel convolutional neural network (CNN) model-based method for detecting salient items in an image. They defined the multi-label classification problem as the salient object detection problem for this reason. Using a CNN that has been trained to predict object shape, their method directly

calculates the salient item's shape. Using hierarchical segmentation maps, they further improve the saliency map that the CNN predicted in order to take advantage of global information like object borders and spatial consistency.[8].

20. The adoption of robotics and self-driving cars is thought to begin with object detection. In this research, they provide light on the function of CNN-based deep learning algorithms for object detection. The article also discusses object detection services and deep learning systems. Additionally covered are benchmarked datasets for object detection and localization that have been made public in international contests. It has been addressed how to find the domains where object detection is useful. Modern deep learning-based object detection methods have been evaluated and contrasted. [9].

CHAPTER III

Methodology

3.1 Basic theory and Algorithms analysis

One kind of Deep Learning neural network architecture that is frequently utilized in computer vision is the convolutional neural network (CNN). The branch of artificial intelligence known as "Computer Vision" gives computers the ability to comprehend and analyze images and other visual input.

1. Convolutional Layer: The convolutional layer, which produces the majority of the network's computations, is the fundamental layer used to build convolutional neural networks. Keep in mind that the number of parameters does not equal the amount of calculation. Compared to a fully connected network of the same size, the convolution operation can effectively reduce the training complexity of the network model as well as the network connection and parameter weights. Standard convolution, transposed convolution, hole convolution, and depth separable convolution are examples of common convolution operations.

2.Activation Layer: Artificial neural networks can be filled with an Activation Function to aid in the network's ability to recognize complex patterns in data. Rectified Linear Units (ReLU), Randomized LeakyReLUs (RReLU), Exponential Linear Units (ELU), and others are examples of common activation functions. Among the most important unsaturated activation functions is the linear rectification function (ReLU). As shown in its mathematical expression is as follows: f(x)=max (0, x).

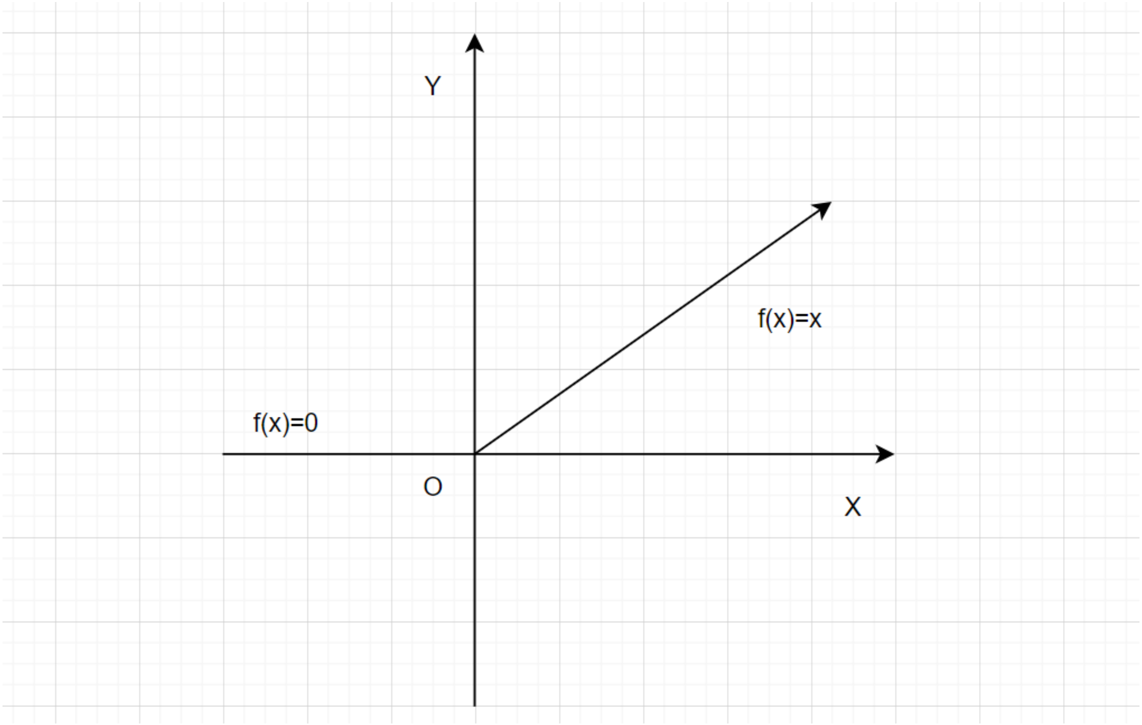

Figure 3.1: ReLU function image

Here figure 3.1 shows the ReLU function image. The ReLU function is a commonly used activation function in artificial neural networks. It's a simple mathematical function that returns the input for any positive input value and returns zero for any negative input value. Mathematically, the ReLU function is defined as:

$$f(x) = max(0, x)$$

3.Pooling Layer: These days, convolutional neural networks frequently use it as one of their constituent parts. In order to reduce overfitting, the amount of data and parameters are compressed by placing the pooling layer between successive convolutional layers. The pooling layer's primary job is to compress images if that's the input type.The pooling layer's primary job when the input is an image is to compress it. By performing collective statistical operations on the special diagnosis at various positions in the local area of the image, the pooling layer can effectively reduce the size of the matrix. This reduces the parameters in the final fully connected layer, speeds up calculation speed, and lessens the excessive sensitivity of the convolutional layer to the image position. The common operations of the pooling layer include the following: max-pooling, average pooling, Spatial Pyramid Pooling, etc.

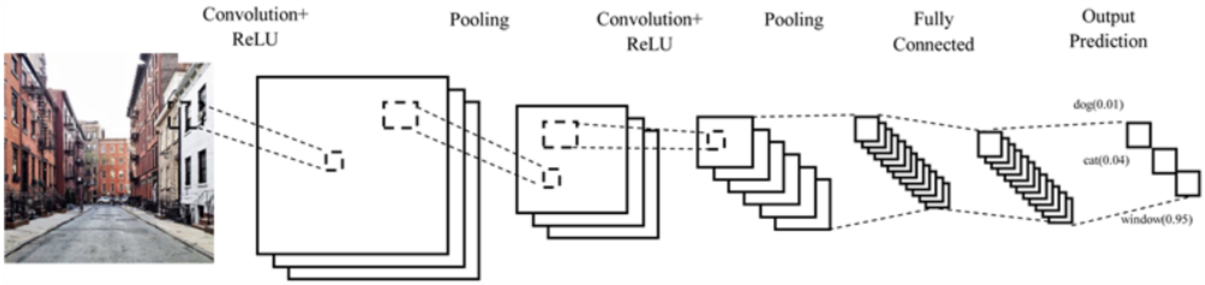

Figure 3.2: Architecture of CNN

Here figure 3.2 shows the Architecture of CNN.Here Input Layer represents the raw input data, typically an image in the case of computer vision applications. The dimensions of the input are usually height, width, and channels (e.g., RGB channels for a color image). Convolutional Layers are the core building blocks of CNNs. They consist of filters (also called kernels) that slide over the input data, convolving to extract features such as edges, textures, and patterns. Convolutional layers help the network learn hierarchical representations. Typically, Rectified Linear Unit (ReLU) activations are applied after convolutional operations to introduce non-linearity to the model, allowing it to learn more complex patterns. Pooling layers follow convolutional layers to reduce the spatial dimensions of the input data and decrease the computational load.After several convolutional and pooling layers, one or more fully connected layers are often added to produce the final output.Before the fully connected layers, the high-dimensional data from the convolutional and pooling layers is flattened into a one-dimensional vector. Batch Normalization is often employed to normalize the inputs of a layer, helping with training stability and accelerating convergence. Dropout is a regularization technique where a random subset of neurons is ignored during training. This helps prevent overfitting.The final layer produces the network's output. For classification tasks, it often involve a softmax activation function to produce class probabilities. The loss function measures the difference between the network's prediction and the actual target values.

Working procedure of CNN

- Import the necessary libraries.
- Set the parameter.
- Defines the kernel.
- Load the image and plot it.
- Reformat the image.
- Apply convolution layer operation and plot the output image.
- Apply activation layer operation and plot the output image.
- Apply pooling layer operation and plot the output image.

Flowchart: Here flowchart CNN has been given below.how it works has been described in this flow diagram.

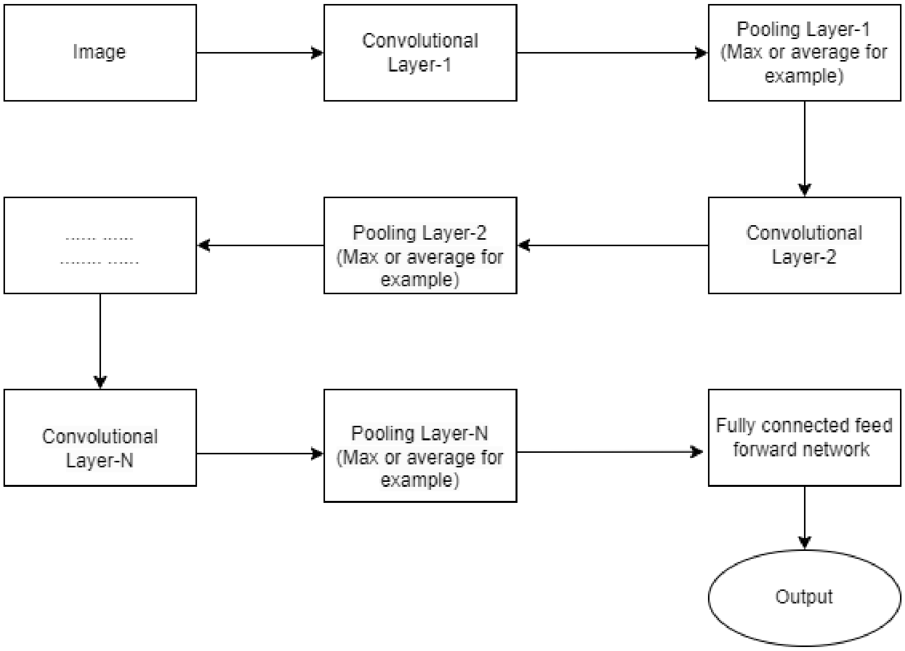
Image
Convolutional Layer-1
Pooling Layer-1 (Max or average for example)
Convolutional Layer-2
Pooling Layer-2 (Max or average for example)
....
....
Convolutional Layer-N
Pooling Layer-N (Max or average for example)
Fully connected feed forward network
Output

Figure 3.3: Flow chart of CNN network

Here figure 3.3 shows the Flow chart of the CNN network. This flowchart describes the way CNN works. It gives us a wider description of each and every process of instruction.

Here are some applications of CNNs:

Image Classification: CNNs are frequently used in image classification applications, where the network is trained to classify images into various groups. This program is widely used in many different domains, from diagnosing medical images to recognizing objects in photos.

Object Detection: CNNs are useful for locating and identifying objects in pictures. CNNs are used in popular object detection frameworks such as Faster R-CNN and YOLO (You Only Look Once). Robotics, autonomous cars, and surveillance systems are a few examples of applications.

D2DET:

A deformable two-stage detector is a type of object detection model used in deep learning for computer vision tasks. It combines two key concepts: the two-stage architecture and deformable convolutional networks.

1.Two-stage architecture: Object detection models typically follow a two-stage approach. In the first stage, they generate region proposals, which are potential bounding boxes that may contain objects. In the second stage, these proposed regions are classified and refined to produce the final object detections.

2..Deformable Convolutional Networks (DCN): Deformable Convolutional Networks are a type of convolutional layer that can adaptively adjust their receptive fields based on the features within an image. This allows them to capture more accurateand flexible information about object shapes, especially when objects are occluded, deformed, or occur at various scales.

Working procedure of D2Det algorithm:

Step 1: Input Image Begin with an input image that you want to perform object detection on.

Step 2: Feature Extraction Pass the input image through a convolutional neural network (CNN) to extract feature maps.These feature maps capture visual information from the image.

Step 3: Region Proposal Network (RPN) The RPN operates on the feature maps and generates a set of anchor boxes (potential object bounding boxes) at various scales and aspect ratios. Each anchor box is associated with a score indicating the likelihood of containing an object. Apply deformable convolutions in the RPN to adaptively adjust the receptive fields and capture more relevant information.

Step 4: Anchor Scoring Assign objectness scores to each anchor box based on how likely they are to contain an object. This is done using a classification layer.

Step 5: Anchor Refinement Predict adjustments (bounding box regressions) for the anchor boxes to better align them with the true object locations. This is done using a regression layer.

Step 6: Anchor Selection Filter the anchor boxes based on their objectness scores, selecting the top-ranked anchors as proposals for further processing. These proposals represent regions likely to contain objects.

Step 7: Region-of-Interest (RoI) Pooling Extract fixed-size feature vectors for each selected proposal by applying RoI pooling or a similar technique on the feature maps. This step ensures that all proposals have consistent input sizes for further processing, regardless of their original sizes.

Step 8: Deformable Convolutional Networks (DCN) Pass the RoI features through deformable convolutional layers, which adaptively adjust their receptive fields. Deformable convolutions help capture information from different position within the RoI, which is particularly beneficial for deformable or occluded objects.

Step 9: Object Classification Use a classifier network to assign a class label to each RoI, determining the type of object contained in the proposal.

Step 10: Bounding Box Regression Apply a regressor network to refine the coordinates of the bounding boxes associated with the RoIs. This step helps adjust the bounding boxes to better fit the precise object locations.

Step 11: Post-processing Perform non-maximum suppression (NMS) to remove redundant or highly overlapping bounding boxes.This ensures that only the most confident detections are retained and eliminates duplicates.

Step 12: Output Provide the final set of detected objects, along with their class labels and refined bounding box coordinates, as the output of the deformable two-stage detector.

The deformable two-stage detector algorithm combines the strengths of deformable convolutional layers with the two-stage object detection approach to achieve improved accuracy in detecting and localizing objects, especially in cases where objects exhibit deformations, occlusions, or variations in scale.

SSD:

SSD operates as a one-shot detector. It predicts the boundary boxes and the classes directly from feature maps in a single pass and does not have a gave region proposal network. SSD adds offsets to default boundary boxes and small convolutional filters to predict object classes in order to increase accuracy. The SSD object detection composes of 2 parts:

1)Extract feature maps.

2)Apply convolution filters to detect objects.

SSD extracts feature maps using VGG16. Next, it makes use of the Conv 4:3 layer to detect objects. As an example, Four object predictions are made for each cell, also known as location. To improve accuracy, SSD can be trained from beginning to end. SSD has better coverage on location, scale, and aspect ratios and makes more predictions.

Working procedure of SSD:

Here's an overview of how an SSD works in deep learning:

1. Input Image: A neural network is used by the SSD model to extract features from an input image of any size. The backbone network is usually a convolutional neural network (CNN) that has been pre-trained, like VGG16 or ResNet. CNN uses several scales to extract features from the picture.

2. Feature Maps: Various-sized feature maps are produced as the image is processed by CNN. Information is captured at different spatial resolutions by these feature maps. In general, higher-level features are captured by the deeper layers of the CNN, whereas lower-level features are captured by the shallower layers.

3. Multi-scale Feature Fusion:

SSD uses convolutional layers with different kernel sizes and moves to obtain feature maps at different scales. These layers are in charge of combining features from the backbone network's various layers. To identify objects of various sizes, feature maps at various scales are utilized.

4. Default Anchor Boxes: A set of default anchor boxes, also called default bounding boxes, with various aspect ratios and scales are defined for every location in the feature maps. To begin object detection, these anchor boxes are used. In order to better fit the real objects in the picture, the SSD model predicts offsets, or deltas, for these anchor boxes.

5. Object Detection Head: After feature extraction and anchor box definition, the SSD network splits into two parallel subnetworks: - Localization Head: This part predicts the offsets (deltas) for each anchor box to adjust their positions and sizes to match the ground-truth objects in the image. - Classification Head: This part predicts the class probabilities for each anchor box, indicating the likelihood that an object of a particular class is present in that box.

6. Non-Maximum Suppression (NMS): Following the receipt of predictions from the localization and classification heads, low-confidence and redundant detections are filtered out using a post-processing technique known as non-maximum suppression. NMS makes sure that as final detections, only the most certain and non-overlapping bounding boxes are kept.

7. Output: An array of bounding boxes with the corresponding class labels and confidence scores is the SSD model's final result. The objects that were found in the input image are represented by these bounding boxes.

YOLO:

YOLO is an algorithm that provides real-time object detection using neural networks. The accuracy and speed of this algorithm make it popular.
The abbreviation YOLO stands for "You Only Look Once." This algorithm (in real-time) finds and recognizes different objects in an image. YOLO employs object detection as a regression problem, resulting in the class probabilities of the identified images. Convolutional neural networks (CNN) are used by the YOLO algorithm to detect objects in real time. As the name implies, the algorithm can detect objects with just one forward propagation via a neural network. This indicates that a single algorithm run is used to predict the entire image. Multiple class probabilities and bounding boxes are simultaneously predicted by the CNN. There are several variations of the YOLO algorithm. Tiny YOLO and YOLOv3 are a couple of the popular ones.

Working procedure of YOLO algorithms:

YOLO algorithm works using the following three techniques:

- Residual blocks
- Bounding box regression
- Intersection Over Union (IOU)

Residual blocks

First, the image is divided into various grids. Each grid has a dimension of S x S. The following image shows how an input image is divided into grids. Bounding box regression A bounding box is an outline that highlights an object in an image. Every bounding box in the image consists of the following attributes:

- Width (bw)
- Height (bh)
- Class (for example, person, car, traffic light, etc.)- This is represented by the letter c.
- Bounding box center (bx,by)

Intersection over union (IOU)

In object detection, the phenomenon known as intersection over union (IOU) characterizes how boxes overlap. YOLO creates an output box that exactly covers the objects by using IOU. The task of predicting the bounding boxes and their confidence scores falls on each grid cell. If the expected and actual bounding boxes match, the IOU is equal to 1. Bounding boxes that are not equal to the actual box are removed by this mechanism.

Combination of the three techniques

The image is first split up into grid cells. B bounding boxes are predicted by each grid cell, along with their confidence scores. To determine each object's class, the cells make predictions about the class probabilities. A car, a dog, and a bicycle are just a few examples of the at least three classes of objects that are visible. A single convolutional neural network is used to make all of the predictions at the same time. The predicted bounding boxes and the actual boxes of the objects are guaranteed to be equal by intersection over union. The effect removes irrelevant bounding boxes that don't match the object's dimensions (height and width). The final detection will be made up of distinct bounding boxes that precisely match the objects. For instance, the yellow bounding box surrounds the bicycle and the pink bounding box covers the car. The blue bounding box has been used to highlight the dog. YOLO algorithm can be applied in the following fields: Autonomous vehicles can utilize the YOLO algorithm to identify nearby objects, including people, cars, and parking signals. In forests, this algorithm is used to identify different kinds of animals. Journalists and wildlife rangers use this kind of detection to recognize animals in photos.

3.2 Proposed System Model

In our proposed topic we used four algorithms. Here is a flowchart of our proposed model.

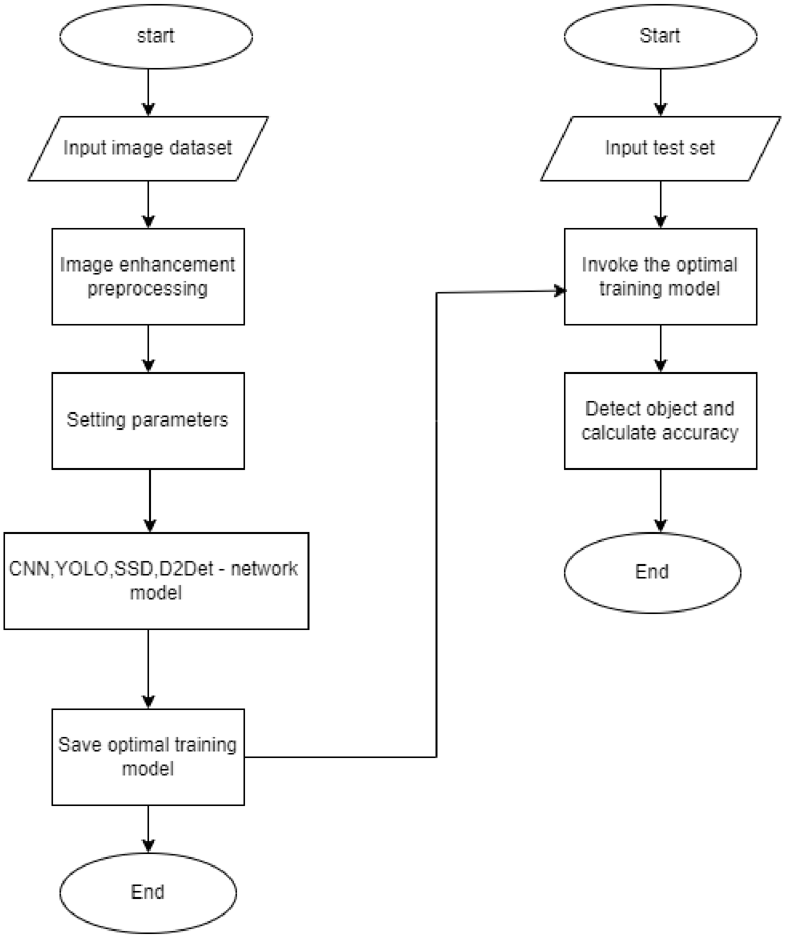
start
Input image dataset
Image enhancement preprocessing
Setting parameters
CNN,YOLO,SSD,D2Det - network model
Save optimal training model
End
Start
Input test set
Invoke the optimal training model
Detect object and calculate accuracy
End

Figure 3.4: Proposed system model

Here figure 3.4 shows the proposed model of our project.

Our system mainly focuses on object detection using 4 algorithms and then also calculate the accuracy of those 4 algorithms. Given flow chart describes our system model in a detailed way. There are two parts of our flow chart. Left part is for preprocessing images/input dataset and getting optimal training model. The right part is for detecting objects and accuracy calculation.

Here are the steps to follow:

1. Start

2. Input image/dataset

3. Image enhancement preprocessing

4. Setting parameters

5. CNN, YOLO,SSD, D2Det network model

6. Save optimal training model

    - Start
    - Input test set
    - Invoke the optimal training model
    - Detect object and calculate accuracy
    - End

7. End

CHAPTER IV

Conclusion and Future Work

4.1 Conclusion

Comparing CNN (Convolutional Neural Networks), YOLO (You Only Look Once), D2Det (Deformable Two-Stage Detector), and SSD (Single Shot MultiBox Detector) algorithms for object detection in deep learning involves considering their strengths, weaknesses, and use cases. Each algorithm has its unique characteristics and is suited for specific scenarios.

4.2 Future work

- Merge Algorithms: We will combine these algorithms and will see whether it gives a better accuracy and result.We will experiment this on various object and conditions.Thus we can make a device that will help others in object detection.
- Real images: We will experiment it on real images.Will see the effect of using this on real images.We will observe the past affect.
- Disease Detection:Wheather a device can be made so that it can say which disease it belongs to.
- Leaf names: Will try to make a device that will recognize a leaf names.
- Road sign: Weahther it can detect road sign from moving objects.

References

[1] Iftikhar Ahmad, Yayun Yang, Yi Yue, Chen Ye, Muhammad Hassan, Xi Cheng, Yunzhi Wu, and Youhua Zhang. Deep learning based detector yolov5 for identifying insect pests. *Applied Sciences*, 12(19):10167, 2022.

[2] Ji Dong Choi and Min Young Kim. A sensor fusion system with thermal infrared camera and lidar for autonomous vehicles and deep learning based object detection. *ICT Express*, 9(2):222–227, 2023.

[3] Tausif Diwan, G Anirudh, and Jitendra V Tembhurne. Object detection using yolo: Challenges, architectural successors, datasets and applications. *multimedia Tools and Applications*, 82(6):9243–9275, 2023.

[4] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[5] Xing Hu, Chun Xie, Zhe Fan, Qianqian Duan, Dawei Zhang, Linhua Jiang, Xian Wei, Danfeng Hong, Guoqiang Li, Xinhua Zeng, et al. Hyperspectral anomaly detection using deep learning: A review. *Remote Sensing*, 14(9):1973, 2022.

[6] Ravpreet Kaur and Sarbjeet Singh. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, 132:103812, 2023.

[7] Huieun Kim, Youngwan Lee, Byeounghak Yim, Eunsoo Park, and Hakil Kim. On-road object detection using deep neural network. In *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4. IEEE, 2016.

[8] Jongpil Kim and Vladimir Pavlovic. A shape-based approach for salient object detection using deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 455–470. Springer, 2016.

[9] Ajeet Ram Pathak, Manjusha Pandey, and Siddharth Rautaray. Application of deep learning for object detection. *Procedia computer science*, 132:1706–1717, 2018.

[10] Junsong Ren and Yi Wang. Overview of object detection algorithms using convolutional neural networks. *Journal of Computer and Communications*, 10(1):115–132, 2022.

[11] Xinyi Shen, Guolong Shi, Huan Ren, and Wu Zhang. Biomimetic vision for zoom object detection based on improved vertical grid number yolo algorithm. *Frontiers in Bioengineering and Biotechnology*, 10:905583, 2022.

[12] Fahima Tabassum, Md Imdadul Islam, Risala Tasin Khan, and M Ruhul Amin. Human face recognition with combination of dwt and machine learning. *Journal of King Saud University-Computer and Information Sciences*, 34(3):546–556, 2022.

[13] Linbo Tang, Wei Tang, Xin Qu, Yuqi Han, Wenzheng Wang, and Baojun Zhao. A scale-aware pyramid network for multi-scale object detection in sar images. *Remote Sensing*, 14(4):973, 2022.

[14] Di Tian, Yi Han, Biyao Wang, Tian Guan, Hengzhi Gu, and Wei Wei. Review of object instance segmentation based on deep learning. *Journal of Electronic Imaging*, 31(4):041205–041205, 2022.

[15] Ayşegül Uçar, Yakup Demir, and Cüneyt Güzeliş. Object recognition and detection with deep learning for autonomous driving applications. *Simulation*, 93(9):759–769, 2017.

[16] Xiongwei Wu, Doyen Sahoo, and Steven CH Hoi. Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64, 2020.

[17] Xuebin Yue, Hengyi Li, Masao Shimizu, Sadao Kawamura, and Lin Meng. Yolo-gd: a deep learning-based object detection algorithm for empty-dish recycling robots. *Machines*, 10(5):294, 2022.

[18] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, 126:103514, 2022.

[19] Yanzhao Zhu and Wei Qi Yan. Traffic sign recognition based on deep learning. *Multimedia Tools and Applications*, 81(13):17779–17791, 2022.