REAL TIME OBJECT DETECTION USING DEEP LEARNING

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

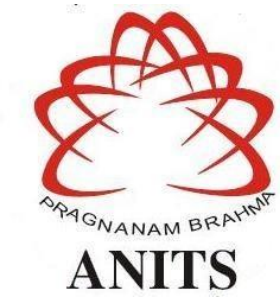
Submitted by

D Pavan (316126512073)
V S Ashlesh Kumar (31612651119)
J.A.S. Sampreeth (316126512139)
K. Sai Naveen (316126512141)

Under the guidance of

Ms. Ch. Padma Sree, M.Tech, (Ph.D)

Assistant Professor, Department of ECE



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES

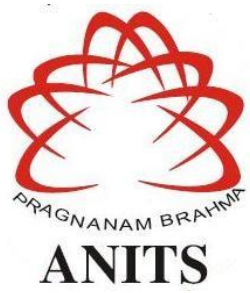
(UGC AUTONOMOUS)

(*Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade*)
Sangivalasa, Bheemilimandal, Visakhapatnam dist.(A.P) 2019-2020

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES

(*Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade*)

Sangivalasa, bheemili mandal, visakhapatnam dist.(A.P)



CERTIFICATE

This is to certify that the project report entitled **"*REAL TIME OBJECT DETECTION USING DEEP LEARNING*"** submitted by ***D Pavan (316126512073), V S Ashlesh Kumar(31612612117), J.A.S. Sampreeth (316126512139), K Sai Naveen (316126512141)*** in partial fulfillment of the requirements for the award of the degree of ***Bachelor of Engineering*** *in* ***Electronics & Communication Engineering*** of Andhra University, Visakhapatnam is a record of bonafide work carried out under my guidance and supervision.

Padma Sree

Project Guide

Ms. Ch Padma Sree

M.Tech,(Ph.D)

Assistant Professor

Department of ECE, ANITS

Head of the Department

Dr. V Rajyalakshmi

M.E, Ph.D., MHRM, MIEEE, MIE, MIETE

Department of E.C.E, ANITS

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our project guide **Ms. Padmasree Challa,** Assistant Professor, M.Tech,(Ph.D) Department of Electronics and Communication Engineering, ANITS, for his/her guidance with unsurpassed knowledge and immense encouragement. We are grateful to **Dr. V. Rajyalakshmi**, Head of the Department, Electronics and Communication Engineering, for providing us with the required facilities for the completion of the project work.

We are very much thankful to the **Principal and Management, ANITS, Sangivalasa,** for their encouragement and cooperation to carry out this work.

We express our thanks to all **teaching faculty** of Department of ECE, whose suggestions during reviews helped us in accomplishment of our project. We would like to thank **all non-teaching staff** of the Department of ECE, ANITS for providing great assistance in accomplishment of our project.

We would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully.

PROJECT STUDENTS:

D Pavan (316126512073)
V S Ashlesh Kumar (31612651119)
J.A.S. Sampreeth (316126512139)
K. Sai Naveen (316126512141)

CONTENTS

ABSTRACT

Real time object detection is a vast, vibrant and complex area of computer vision. Ifthere is a single object to be detected in an image, it is known as Image Localization and ifthere are multiple objects in an image, then it is Object Detection. This detects the semanticobjects of a class in digital images and videos. The applications of real time object detectioninclude tracking objects, video surveillance, pedestrian detection, people counting, self-driving cars, face detection, ball tracking in sports and many more. Convolution NeuralNetworks is a representative tool of Deep learning to detect objects using OpenCV(Opensource Computer Vision), which is a library of programming functions mainly aimed at realtime computer vision.

Keywords: Computer vision, Deep Learning, Convolution Neural Networks.

Team Members:

D Pavan (316126512079)

V S Ashlesh (316126512073)

J.A.S.Sampreeth (316126512139)

K Sai Naveen (31616512141)

Project Guide:

Ms. Ch. Padma Sree, Asst. Prof., ECE

CHAPTER 1

INTRODUCTION

Project Objective:

The motive of object detection is to recognize and locate all known objects in a scene. Preferably in 3D space, recovering pose of objects in 3D is very important for robotic control systems.

Imparting intelligence to machines and making robots more and more autonomous and independent has been a sustaining technological dream for the mankind. It is our dream to let the robots take on tedious, boring, or dangerous work so that we can commit our time to more creative tasks. Unfortunately, the intelligent part seems to be still lagging behind. In real life, to achieve this goal, besides hardware development, we need the software that can enable robot the intelligence to do the work and act independently. One of the crucial components regarding this is vision, apart from other types of intelligences such as learning and cognitive thinking. A robot cannot be too intelligent if it cannot see and adapt to a dynamic environment.

The searching or recognition process in real time scenario is very difficult. So far, no effective solution has been found for this problem. Despite a lot of research in this area, the methods developed so far are not efficient, require long training time, are not suitable for real time application, and are not scalable to large number of classes. Object detection is relatively simpler if the machine is looking for detecting one particular object. However, recognizing all the objects inherently requires the skill to differentiate one object from the other, though they may be of same type. Such problem is very difficult for machines, if they do not know about the various possibilities of objects.

Motivation:

Blind people do lead a normal life with their own style of doing things. But, they definitely face troubles due to inaccessible infrastructure and social challenges. The biggest challenge for a blind person, especially the one with the complete loss of vision, is to navigate around places. Obviously, blind people roam easily around their house without any help because they know the position of everything in the house. Blind people have a tough time finding objects around them. . So we decided to make a REAL TIME OBJECT DETECTION System. We are interested in this project after we went through few papers in this area. As a result we are highly motivated to develop a system that recognizes objects in the real time environment

CHAPTER 2

OBJECT DETECTION

INTRODUCTION TO OBJECT DETECTION

Object Detection is the process of finding and recognizing real-world object instances such as car, bike, TV, flowers, and humans out of an images or videos. An object detection technique lets you understand the details of an image or a video as it allows for the recognition, localization, and detection of multiple objects within an image.

It is usually utilized in applications like image retrieval, security, surveillance, and advanced driver assistance systems (ADAS).Object Detection is done through many ways:

- Feature Based Object Detection
- Viola Jones Object Detection
- SVM Classifications with HOG Features
- Deep Learning Object Detection

Object detection from a video in video surveillance applications is the major task these days. Object detection technique is used to identify required objects in video sequences and to cluster pixels of these objects.

The detection of an object in video sequence plays a major role in several applications specifically as video surveillance applications.

Object detection in a video stream can be done by processes like pre-processing, segmentation, foreground and background extraction, feature extraction.

Humans can easily detect and identify objects present in an image. The human visual system is fast and accurate and can perform complex tasks like identifying multiple objects with little conscious thought. With the availability of large amounts of data, faster GPUs, and better algorithms, we can now easily train computers to detect and classify multiple objects within an image with high accuracy.

DIGITAL IMAGE PROCESSING

Computerized picture preparing is a range portrayed by the requirement for broad test work to build up the practicality of proposed answers for a given issue. A critical trademark hidden the plan of picture preparing frameworks is the huge level of testing and experimentation that

Typically is required before touching base at a satisfactory arrangement. This trademark infors that the capacity to plan approaches and rapidly model hopeful arrangements by and large assumes a noteworthy part in diminishing the cost and time required to land at a suitableframework execution.

WHAT IS DIP?

A picture might be characterized as a two-dimensional capacity f(x, y), where x, y are spatial directions, and the adequacy off at any combine of directions (x, y) is known as the power or dark level of the picture by then. Whenever x, y and the abundance estimation of are all limited discrete amounts, we call the picture a computerized picture. The field of DIP alludes to preparing advanced picture by methods for computerized PC. Advanced picture is made out of a limited number of components, each of which has a specific area and esteem. The components are called pixels.

Vision is the most progressive of our sensor, so it is not amazing that picture play the absolute most imperative part in human observation. Be that as it may, dissimilar to people, who are constrained to the visual band of the EM range imaging machines cover practically the whole EM range, going from gamma to radio waves. They can work likewise on pictures produced by sources that people are not acclimated to partner with picture.

There is no broad understanding among creators in regards to where picture handling stops and other related territories, for example, picture examination and PC vision begin. Now and then a qualification is made by characterizing picture handling as a teach in which both the info and yield at a procedure are pictures. This is constraining and to some degree manufactured limit. The range of picture investigation is in the middle of picture preparing and PC vision.

There are no obvious limits in the continuum from picture handling toward one side to finish vision at the other. In any case, one helpful worldview is to consider three sorts of mechanized procedures in this continuum: low, mid and abnormal state forms. Low-level process includes primitive operations, for example, picture preparing to decrease commotion differentiate upgrade and picture honing. A low-level process is described by the way that both its sources of info and yields are pictures.

Mid-level process on pictures includes assignments, for example, division, depiction of that

Question diminish them to a frame reasonable for PC handling and characterization of individualarticles

A mid-level process is portrayed by the way that its sources of info by and large are pictures however its yields are properties removed from those pictures. At long last more clevated amount handling includes "Understanding an outlet of perceived items, as in picture examination and at the farthest end of the continuum playing out the intellectual capacities typically connected with human vision. Advanced picture handling, as effectively characterized is utilized effectively in a wide scope of regions of outstanding social and monetary esteem.

WHAT IS AN IMAGE?

A picture is spoken to as a two dimensional capacity f(x, y) where x and y are spatial co-ordinates and the adequacy of "T" at any match of directions (x, y) is known as the power of the picture by then.

Fig. 1.1 digital image

Processing on image:

Processing on image can be of three types They are low-level, mid-level, high level.

Low-level Processing:

- Preprocessing to remove noise.
- Contrast enhancement.
- Image sharpening.

Medium Level Processing :

- Segmentation.
- Edge detection
- Object extraction.

High Level Processing:

- Image analysis
- Scene interpretation

Why Image Processing?

Since the digital image is invisible, it must be prepared for viewing on one or more output device(laser printer, monitor at).The digital image can be optimized for the application by enhancing the appearance of the structures within it.

There are three of image processing used. They are

- Image to Image transformation
- Image to Information transformations
- Information to Image transformations

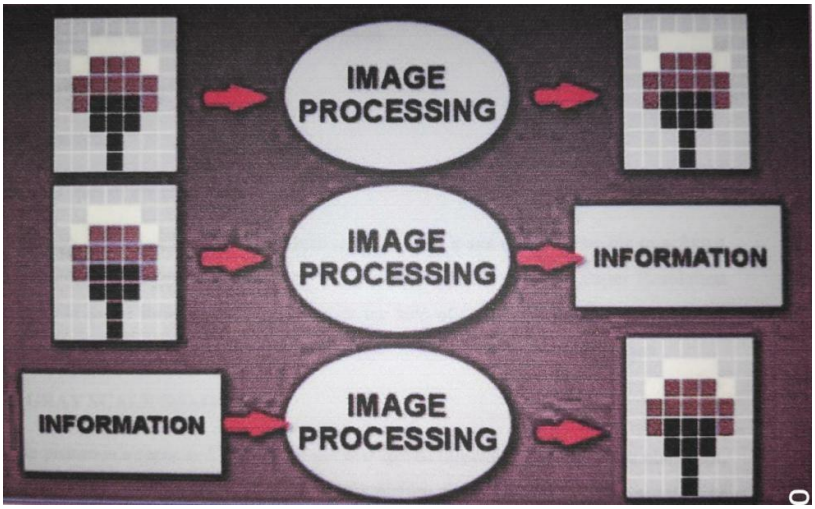


Fig. 1.2 Types of Image Processing

Pixel :

Pixel is the smallest element of an image. Each pixel correspond to any one value. In an 8-bit gray scale image, the value of the pixel between 0 and 255.Each pixel store a value proportional to the light intensity at that particular location. It is indicated in either Pixels per inch or Dots per inch.

Resolution :

The resolution can be defined in many ways. Such as pixel resolution, spatial resolution, temporal resolution, spectral resolution.In pixel resolution, the term resolution refers to the total number of count of pixels in an digital image. For example, If an image has M rows and N columns, then its resolution can be defined as MX N. Higher is the pixel resolution, the higher is the quality of the image.

Resolution of an image is of generally two types.

• Low Resolution image

• High Resolution

Since high resolution is not a cost effective process It is not always possible to achieve high resolution images with low cost. Hence it is desirable Imaging. In Super Resolution imaging, with the help of certain methods and algorithms we can be able to produce high resolution images from the low resolution image from the low resolution images.

GRAY SCALE IMAGE

A gray scale picture is a capacity I (xylem) of the two spatial directions of the picture plane. I(x,y) is the force of the picture force of picture at the point (x, y) on the picture plane. I (xylem) take non-negative expect the picture is limited by a rectangle

COLOR IMAGE

It can be spoken to by three capacities, R (xylem) for red, G (xylem) for green and B (xylem) for blue. A picture might be persistent as for the x and y facilitates and furthermore in adequacy. Changing over

such a picture to advanced shape requires that the directions and the adequacy to be digitized. Digitizing the facilitate's esteems is called inspecting. Digitizing the adequacy esteems is called quantization.

RELATED TECHNOLOGY:

R-CNN

R-CNN is a progressive visual object detection system that combines bottom-up region proposals with rich options computed by a convolution neural network.

R-CNN uses region proposal ways to initial generate potential bounding boxes in a picture and then run a classifier on these proposed boxes.

SINGLE SIZE MULTI BOX DETECTOR

SSD discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At the time of prediction the network generates scores for the presence of each object category in each default box and generates adjustments to the box to better match the object shape.

Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes.

ALEXNET

AlexNet is a convolutional neural Network used for classification which has 5 Convolutional layers, 3 fullyconnected layers and 1 softmax layer with 1000 outputs for classification as his architecture.

YOLO

YOLO is real-time object detection. It applies one neural network to the complete image dividing the image into regions and predicts bounding boxes and possibilities for every region.

Predicted probabilities are the basis on which these bounding boxes are weighted. A single neural network predicts bounding boxes and class possibilities directly from full pictures in one evaluation. Since the full detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

VGG

VGG network is another convolution neural network architecture used for image classification.

MOBILENETS

To build lightweight deep neural networks MobileNets are used. It is based on a streamlined architecture that uses depth-wise separable convolutions. MobileNet uses 3×3 depth-wise separable convolutions that uses between 8 times less computation than standard convolution at solely a little reduction accuracy. Applications and use cases including object detection, fine grain classification, face attributes and large scale-localization.

TENSOR FLOW

Tensor flow is an open source software library for high performance numerical computation. It allows simple deployment of computation across a range of platforms (CPUs, GPUs, TPUs) due to its versatile design also from desktops to clusters of servers to mobile and edge devices. Tensor flow was designed and developed by researchers and engineers from the Google Brain team at intervals Google's AI organization, it comes with robust support for machine learning and deep learning and the versatile numerical computation core is used across several alternative scientific domains.

To construct, train and deploy Object Detection Models TensorFlow is used that makes it easy and also it provides a collection of Detection Models pre-trained on the COCO dataset, the Kitti dataset, and the Open Images dataset. One among the numerous Detection Models is that the combination of Single Shot Detector (SSDs) and Mobile Nets architecture that is quick, efficient and doesn't need huge computational capability to accomplish the object Detection.

APPLICATION OF OBJECT DETECTION

The major applications of Object Detection are:

FACIAL RECOGNITION

"Deep Face" is a deep learning facial recognition system developed to identify human faces in a digital image. Designed and developed by a group of researchers in Facebook. Google also has its own facial recognition system in Google Photos, which automatically separates all the photos according to the person in the image.

There are various components involved in Facial Recognition or authors could say it focuses on various aspects like the eyes, nose, mouth and the eyebrows for recognizing a faces.

PEOPLE COUNTING

People counting is also a part of object detection which can be used for various purposes like finding person or a criminal; it is used for analysing store performance or statistics of crowd during festivals. This process is considered a difficult one as people move out of the frame quickly.

INDUSTRIAL QUALITY CHECK

Object detection also plays an important role in industrial processes to identify or recognize products. Finding a particular object through visual examination could be a basic task that's involved in multiple industrial processes like sorting, inventory management, machining, quality management, packaging and so on. Inventory management can be terribly tough as things are hard to trace in real time. Automatic object counting and localization permits improving inventory accuracy.

SELF DRIVING CARS

Self-driving is the future most promising technology to be used, but the working behind can be very complex as it combines a variety of techniques to perceive their surroundings, including radar, laser light, GPS, odometer, and computer vision. Advanced control systems interpret sensory info to allow navigation methods to work, as well as obstacles and it. This is a big step towards Driverless cars as it happens at very fast speed.

SECURITY

Object Detection plays a vital role in the field of Security; it takes part in major fields such as face ID of Apple or the retina scan used in all the sci-fi movies. Government also widely use this application to access the security feed and match it with their existing database to find any criminals or to detecting objects like car number involved in criminal activities. The applications are limitless.

OBJECT DETECTION WORKFLOW AND FEATURE EXTRACTION

Every Object Detection Algorithm works on the same principle and it’s just the working that differs from others. They focus on extracting features from the images that are given as the input at hands and then it uses these features to determine the class of the image.

CHAPTER 3

DEEP

LEARNING

INTRODUCTION

Deep learning is a machine learning technique. It teaches a computer to filter inputs through layers to learn how to predict and classify information. Observations can be in the form of images, text, or sound. The inspiration for deep learning is the way that the human brain filters information. Its purpose is to mimic how the human brain works to create some real magic. In the human brain, there are about 100 billion neurons. Each neuron connects to about 100,000 of its neighbors. We're kind of recreating that, but in a way and at a level that works for machines. In our brains, a neuron has a body, dendrites, and an axon. The signal from one neuron travels down the axon and transfers to the dendrites of the next neuron. That connection where the signal passes is called a synapse. Neurons by themselves are kind of useless. But when you have lots of them, they work together to create some serious magic. That's the idea behind a deep learning algorithm! You get input from observation and you put your input into one layer. That layer creates an output which in turn becomes the input for the next layer, and so on. This happens over and over until your final output signal! The neuron (**node**) gets a signal or signals (**input values**), which pass through the neuron. That neuron delivers the **output signal**.

Think of the input layer as your senses: the things you see, smell, and feel, for example. These are independent variables for one single observation. This information is broken down into numbers and the bits of binary data that a computer can use. You'll need to either standardize or normalize these variables so that they're within the same range. They use many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output of the previous layer for its input. What they learn forms a hierarchy of concepts. In this hierarchy, each level learns to transform its input data into a more and more abstract and composite representation. That means that for an image, for example, the input might be a matrix of pixels. The first layer might encode the edges and compose the pixels. The next layer might compose an arrangement of edges. The next layer might encode a nose and eyes. The next layer might recognize that the image contains a face, and so on.

What happens inside the neuron?

The input node takes in information in a numerical form. The information is presented as an activation value where each node is given a number. The higher the number, the greater the activation. Based on the connection strength (weights) and transfer function, the activation value passes to the next node. Each of the nodes sums the activation values that it receives (it calculates the **weighted sum**) and modifies that sum based on its transfer function. Next, it applies an activation function. An activation

function is a function that's applied to this particular neuron. From that, the neuron understands if it needs to pass along a signal or not.

Each of the synapses gets assigned weights, which are crucial to **Artificial Neural Networks** (ANNs). Weights are how ANNs learn. By adjusting the weights, the ANN decides to what extent signals get passed along. When you're training your network, you're deciding how the weights are adjusted.

The activation runs through the network until it reaches the output nodes. The output nodes then give us the information in a way that we can understand. Your network will use a cost function to compare the output and the actual expected output. The model performance is evaluated by the cost function. It's expressed as the difference between the actual value and the predicted value.

There are many different cost functions you can use, you're looking at what the error you have in your network is. You're working to minimize loss function. (In essence, the lower the loss function, the closer it is to your desired output). The information goes back, and the neural network begins to learn with the goal of minimizing the cost function by tweaking the weights. This process is called **backpropagation**.

In **forward propagation**, information is entered into the input layer and propagates forward through the network to get our output values. We compare the values to our expected results. Next, we calculate the errors and propagate the info backward. This allows us to train the network and update the weights. (Backpropagation allows us to adjust all the weights simultaneously.) During this process, because of the way the algorithm is structured, you're able to adjust all of the weights simultaneously. This allows you to see which part of the error each of your weights in the neural network is responsible for.

When you've adjusted the weights to the optimal level, you're ready to proceed to the testing phase!

How does an artificial neural network learn?

There are two different approaches to get a program to do what you want. First, there's the specifically guided and hard-programmed approach. You tell the program exactly what you want it to do. Then there are **neural networks**. In neural networks, you tell your network the inputs and what you want for the outputs, and then you let it learn on its own.

By allowing the network to learn on its own, you can avoid the necessity of entering in all of the rules. You can create the architecture and then let it go and learn. Once it's trained up, you can give it a new image and it will be able to distinguish output.

Feedforward and feedback networks

A **feedforward** network is a network that contains inputs, outputs, and hidden layers. The signals can only travel in one direction (forward). Input data passes into a layer where calculations are performed. Each processing element computes based upon the weighted sum of its inputs. The new values become the new input values that feed the next layer (feed-forward). This continues through all the layers and determines the output. Feedforward networks are often used in, for example, data mining.

A **feedback network** (for example, a recurrent neural network) has feedback paths. This means that they can have signals traveling in both directions using loops. All possible connections between neurons are allowed. Since loops are present in this type of network, it becomes a non-linear dynamic system which changes continuously until it reaches a state of equilibrium. Feedback networks are often used in optimization problems where the network looks for the best arrangement of interconnected factors.

Weighted Sum

Inputs to a neuron can either be features from a training set or outputs from the neurons of a previous layer. Each connection between two neurons has a unique synapse with a unique weight attached. If you want to get from one neuron to the next, you have to travel along the synapse and pay the "toll" (weight). The neuron then applies an activation function to the sum of the weighted inputs from each incoming synapse. It passes the result on to all the neurons in the next layer. When we talk about updating weights in a network, we're talking about adjusting the weights on these synapses.

A neuron's input is the sum of weighted outputs from all the neurons in the previous layer. Each input is multiplied by the weight associated with the synapse connecting the input to the current neuron. If there are 3 inputs or neurons in the previous layer, each neuron in the current layer will have 3 distinct weights: one for each synapse.

In a nutshell, the activation function of a node defines the output of that node.

The activation function (or transfer function) translates the input signals to output signals. It maps the output values on a range like 0 to 1 or -1 to 1. It's an abstraction that represents the rate of action potential firing in the cell. It's a number that represents the likelihood that the cell will fire. At it's simplest, the function is binary: **yes** (the neuron fires) or **no** (the neuron doesn't fire). The output can be either 0 or 1 (on/off or yes/no), or it can be anywhere in a range. If you were using a function that maps a range between 0 and 1 to determine the likelihood that an image is a cat, for example, an output of 0.9 would show a 90% probability that your image is, in fact, a cat.

Activation function

In a nutshell, the activation function of a node defines the output of that node.

The activation function (or transfer function) translates the input signals to output signals. It maps the output values on a range like 0 to 1 or -1 to 1. It's an abstraction that represents the rate of action potential firing in the cell. It's a number that represents the likelihood that the cell will fire. At it's simplest, the function is binary: **yes** (the neuron fires) or **no** (the neuron doesn't fire). The output can be either 0 or 1 (on/off or yes/no), or it can be anywhere in a range.

What options do we have? There are many activation functions, but these are the four very common ones:

Thresholdfunction

This is a step function. If the summed value of the input reaches a certain threshold the function passes on 0. If it's equal to or more than zero, then it would pass on 1. It's a very rigid, straightforward, yes or no function.

Sigmoid function

This function is used in logistic regression. Unlike the threshold function, it's a smooth, gradual progression from 0 to 1. It's useful in the output layer and is used heavily for linear regression.

Hyperbolic Tangent Function

This function is very similar to the sigmoid function. But unlike the sigmoid function which goes from 0 to 1, the value goes below zero, from -1 to 1. Even though this isn't a lot like what happens in a brain, this function gives better results when it comes to training neural networks. Neural networks sometimes get "stuck" during training with the sigmoid function. This happens when there's a lot of strongly negative input that keeps the output near zero, which messes with the learning process.

Rectifier function

This might be the most popular activation function in the universe of neural networks. It's the most efficient and biologically plausible. Even though it has a kink, it's smooth and gradual after the kink at 0. This means, for example, that your output would be either "no" or a percentage of "yes." This function doesn't require normalization or other complicated calculations.

The field of artificial intelligence is essential when machines can do tasks that typically require human intelligence. It comes under the layer of machine learning, where machines can acquire skills and learn from past experience without any involvement of human. Deep learning comes under machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. The concept of deep learning is based on humans' experiences; the deep learning algorithm would perform a task continuously so that it can improve the outcome. Neural networks have various (deep) layers that enable learning. Any drawback that needs "thought" to work out could be a drawback deep learning can learn to unravel.

CHAPTER 4

CONVOULUTION NEURAL NETWORKS

INTRODUCTION TO CONVOLUTIONAL NEURAL NETWORKS (CNN)

Artificial Neural Networks

The idea of ANNs is based on the belief that working of human brain by making the right connections, can be imitated using silicon and wires as living neurons and dendrites.

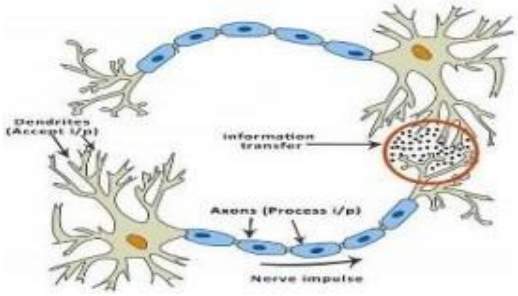


Fig: 4.1

The human brain is composed of 86 billion nerve cells called neurons. They are connected to other thousand cells by Axons. Stimuli from external environment or inputs from sensory organs are accepted by dendrites. These inputs create electric impulses, which quickly travel through the neural network. A neuron can then send the message to other neuron to handle the issue or does not send it forward.

ANNs are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation or node value. Each link is associated with weight. ANNs are capable of learning, which takes place by altering weight values.

Neural network:

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problem. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear

combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be - 1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks,

which can derive conclusions from a complex and seemingly unrelated set of information.

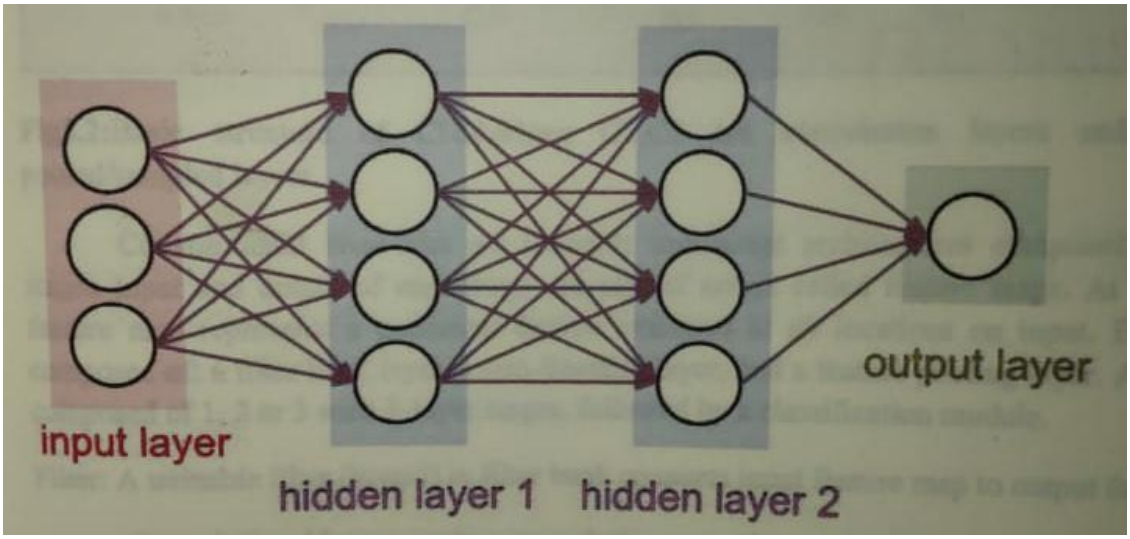


Fig. 4.2 A simple neural network

A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship.

CONVOLUTIONAL NEURAL NETWORKS:

Convolutional Neural Networks are very similar to ordinary Neural Networks from the previous chapter: they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity.

Convolutional Neural Networks (CNNs) are analogous to traditional ANNs in that they are comprised of neurons that self-optimise through learning. Each neuron will still receive an input and perform a operation (such as a scalar product followed by a non-linear function) - the basis of countless ANNs. From the input raw image vectors to the final output of the class score, the entire of the network will still express a single

perceptive score function (the weight). The last layer will contain loss functions associated with the classes, and all of the regular tips and tricks developed for traditional ANNs still apply.

The only notable difference between CNNs and traditional ANNs is that CNNs are primarily used in the field of pattern recognition within images. This allows us to encode image-specific features into the architecture, making the network more suited for image-focused tasks - whilst further reducing the parameters required to set up the model. One of the largest limitations of traditional forms of ANN is that they tend to struggle with the computational complexity required to compute image data. Common machine learning benchmarking datasets such as the MNIST database of handwritten digits are suitable for most forms of ANN, due to its relatively small image dimensionality of just 28 × 28. With this dataset a single neuron in the first hidden layer will contain 784 weights (28×28×1 where 1 bare in mind that MNIST is normalised to just black and white values), which is manageable for most forms of ANN. If you consider a more substantial coloured image input of 64 × 64, the number of weights on just a single neuron of the first layer increases substantially to 12, 288. Also take into account that to deal with this scale of input, the network will also need to be a lot larger than one used to classify colour-normalised MNIST digits, then you will understand the drawbacks of using such models.

CNN ARCHITECTURE:

CNNs are feedforward networks in that information flow takes place in one direction only, from their inputs to their outputs. Just as artificial neural networks (ANN) are biologically inspired, so are CNNs. The visual cortex in the brain, which consists of alternating layers of simple and complex cells (Hubel & Wiesel, 1959, 1962), motivates their architecture.

CNN architectures come in several variations; however, in general, they consist of convolutional and pooling (or subsampling) layers, which are grouped into modules. Either one or more fully connected layers, as in a standard feedforward neural network, follow these modules. Modules are often stacked on top of each other to form a deep model. It illustrates typical CNN architecture for a toy image classification task. An image is input directly to the network, and this is followed by several stages of convolution and pooling. Thereafter, representations from these operations feed one or more fully connected layers.

Finally, the last fully connected layer outputs the class label. Despite this being the most popular base architecture found in the literature, several architecture changes have been proposed in recent years with the objective of improving image classification accuracy or reducing computation costs. Although for the remainder of this section, we merely fleetingly introduce standard CNN architecture.

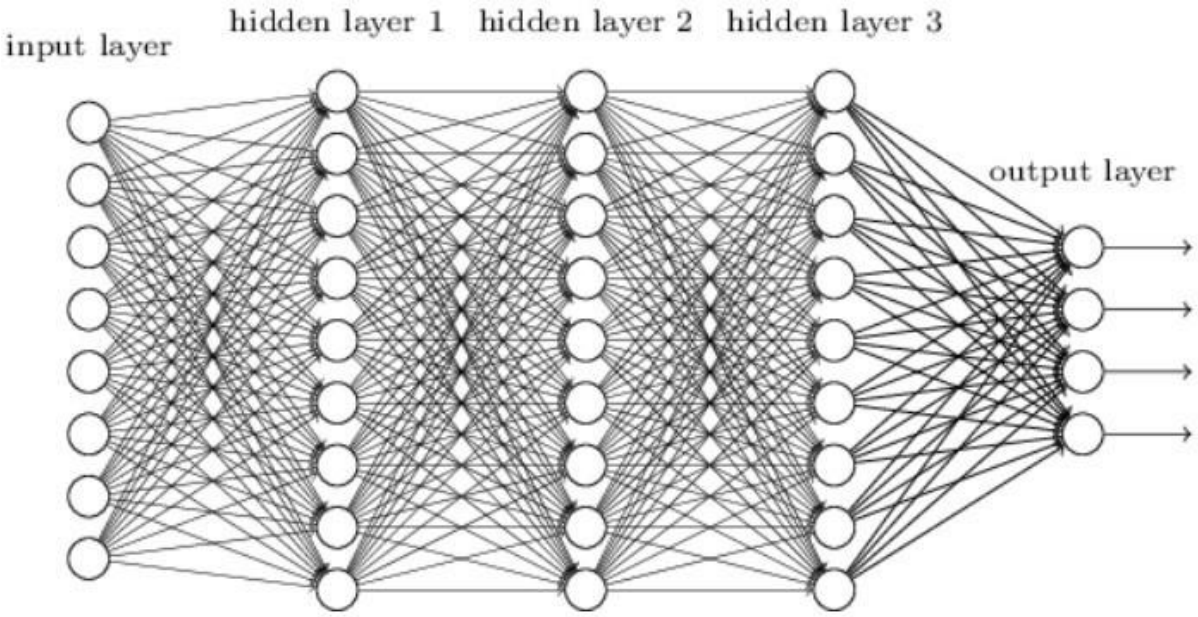


Fig: 4.3

OVERALL ARCHITECTURE:

CNNs are comprised of three types of layers. These are convolutional layers, pooling layers and fully-connected layers. When these layers are stacked, a CNN architecture has been formed. A simplified CNN architecture for MNIST classification is illustrated in Figure 2. input 0 9 convolution w/ReLu pooling output fully-connected w/ ReLu fully-connected ... Fig. 2: An simple CNN architecture, comprised of just five layers The basic functionality of the example CNN above can be broken down into four key areas. 1. As found in other forms of ANN, the input layer will hold the pixel values of the image. 2. The convolutional layer will determine the output of neurons of which are connected to local regions of the input through the calculation of the scalar product between their weights and the region connected to the input volume. The rectified linear unit (commonly shortened to ReLu) aims to apply an 'elementwise' activation function such as sigmoid to the output of the activation produced by the previous layer. 3. The pooling layer will then simply perform downsampling along the spatial dimensionality of the given input, further reducing the number of parameters within that activation. 4. The fully-connected layers will then perform the same duties found in standard ANNs and attempt to produce class scores from the activations, to be used for classification. It is also suggested that ReLu may be used between these layers, as to improve performance. Through this simple method of transformation, CNNs are able to transform the original input layer by layer using convolutional and downsampling techniques to produce class scores for classification and regression purposes. However, it is important to note that simply

understanding the overall architecture of a CNN architecture will not suffice. The creation and optimisation of these models can take quite some time, and can be quite confusing. We will now explore in detail the individual layers, detailing their hyperparameters and connectivities.

CONVOLUTIONAL LAYERS:

The convolutional layers serve as feature extractors, and thus they learn the feature representations of their input images. The neurons in the convolutional layers are arranged into feature maps. Each neuron in a feature map has a receptive field, which is connected to a neighborhood of neurons in the previous layer via a set of trainable weights, sometimes referred to as a filter bank. Inputs are convolved with the learned weights in order to compute a new feature map, and the convolved results are sent through a nonlinear activation function.

All neurons within a feature map have weights that are constrained to be equal; however, different feature maps within the same convolutional layer have different weights so that several features can be extracted at each location.

As the name implies, the convolutional layer plays a vital role in how CNNs operate. The layers parameters focus around the use of learnable kernels.

These kernels are usually small in spatial dimensionality, but spreads along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map. These activation maps can be visualised.

As we glide through the input, the scalar product is calculated for each value in that kernel. From this the network will learn kernels that 'fire' when they see a specific feature at a given spatial position ofthe input. These are commonly known as activations.

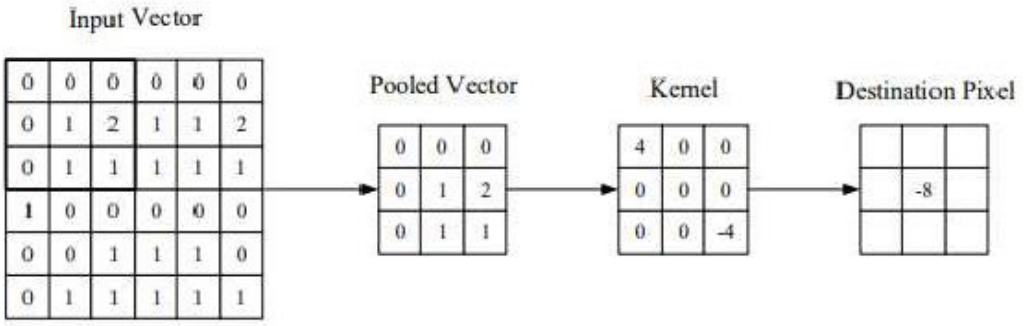


Fig: 4.4 Visual representation of a convolutional layerz

The centre element of the kernel is placed over the input vector, of which is then calculated and replaced with a weighted sum of itself and any nearby pixels.

Every kernel will have a corresponding activation map, of which will be stacked along the depth dimension to form the full output volume from the convolutional layer.

As we alluded to earlier, training ANNs on inputs such as images results in models of which are too big to train effectively. This comes down to the fullyconnected manner of stan ard ANN neurons, so to mitigate against this every neuron in a convolutional layer is only connected to small region of the input volume. The dimensionality of this region is commonly referred to as the receptive field size of the neuron. The magnitude of the connectivity through the depth is nearly always equal to the depth of the input.

For example, if the input to the network is an image of size 64 × 64 × 3 (aRGBcoloured image with a dimensionality of 64 × 64) and we set the receptive field size as 6 × 6, we would have a total of 108 weights on each neuron within the convolutional layer. (6 × 6 × 3 where 3 is the magnitude of connectivity across the depth of the volume) To put this into perspective, a standard neuron seen in other forms of ANN would contain 12, 288 weights each.

Convolutional layers are also able to significantly reduce the complexity of the model through the optimisation of its output. These are optimised through three hyperparameters, the depth, the stride and setting zero-padding.

The depth of the output volume produced by the convolutional layers can be manually set through the number of neurons within the layer to a the same region of the input. This can be seen with other forms of ANNs, where the all of the neurons in the hidden layer are directly connected to every single neuron beforehand. Reducing this hyperparameter can significantly minimise the total number of neurons of the network, but it can also significantly reduce the pattern recognition capabilities of the model.

We are also able to define the stride in which we set the depth around the spatial dimensionality of the input in order to place the receptive field. For example if we were to set a stride as 1, then we would have a heavily overlapped receptive field producing extremely large activations. Alternatively, setting the stride to a greater number will reduce the amount of overlapping and produce an output of lower spatial dimensions.

Zero-padding is the simple process of padding the border of the input, and is an effective method to give further control as to the dimensionality of the output volumes.

It is important to understand that through using these techniques, we will alter the spatial dimensionality of the convolutional layers output.

Despite our best efforts so far we will still find that our models are still enormous if we use an image input of any real dimensionality. However, methods have been developed as to greatly curtail the overall number of parameters within the convolutional layer.

Parameter sharing works on the assumption that if one region feature is useful to compute at a set spatial region, then it is likely to be useful in another region. If we constrain each individual activation map within the output volume to the same weights and bias, then we will see a massive reduction in the number of parameters being produced by the convolutional layer.

As a result of this as the backpropagation stage occurs, each neuron in the output will represent the overall gradient of which can be totalled across the depth - thus only updating a single set of weights, as opposed to every single one.

Pooling Layers

The purpose of the pooling layers is to reduce the spatial resolution of the feature maps and thus achieve spatial invariance to input distortions and translations. Initially, it was common practice to use average pooling aggregation layers to propagate the average of all the input values, of a small neighbourhood of an image to the next layer. However, in more recent models, max pooling aggregation layers propagate the maximum value within a receptive field to the next layer.

Pooling layers aim to gradually reduce the dimensionality of the representation, and thus further reduce the number of parameters and the computational complexity of the model.

The pooling layer operates over each activation map in the input, and scales its dimensionality using the “MAX” function. In most CNNs, these come in the form of max-pooling layers with kernels of

a dimensionality of 2 × 2 applied with a stride of 2 along the spatial dimensions of the input. This scales the activation map down to 25% of the original size - whilst maintaining the depth volume to its standard size.

Due to the destructive nature of the pooling layer, there are only two generally observed methods of max-pooling. Usually, the stride and filters of the pooling layers are both set to 2 × 2, which will allow the layer to extend through the entirety of the spatial dimensionality of the input. Furthermore overlapping pooling may be utilised, where the stride is set to 2 with a kernel size set to 3. Due to the destructive nature of pooling, having a kernel size above 3 will usually greatly decrease the performance of the model.

It is also important to understand that beyond max-pooling, CNN architectures may contain general-pooling. General pooling layers are comprised of pooling neurons that are able to perform a multitude of common operations including L1/L2-normalisation, and average pooling. However, this tutorial will primarily focus on the use of max-pooling.

Fully Connected Layers

Several convolutional and pooling layers are usually stacked on top of each other to extract more abstract feature representations in moving through the network. The fully connected layers that follow these layers interpret these feature representations and perform the function of high-level reasoning. . For classification problems, it is standard to use the softmax operator on top of a DCNN. While early success was enjoyed by using radial basis functions (RBFs), as the classifier on top of the convolutional towers found that replacing the softmax operator with a support vector machine (SVM) leads to improved classification accuracy.

The fully-connected layer contains neurons of which are directly connected to the neurons in the two adjacent layers, without being connected to any layers within them. This is analogous to way that neurons are arranged in traditional forms of ANN.

Despite the relatively small number of layers required to form a CNN, there is no set way of formulating a CNN architecture. That being said, it would be idiotic to simply throw a few of layers together and expect it to work. Through reading of related literature it is obvious that much like other forms of ANNs, CNNs tend to follow a common architecture. This common architecture is illustrated in Figure 2, where convolutional layers are stacked, followed by pooling layers in a repeated manner before feeding forward to fully-connected layers.

Convolutional Neural Networks differ to other forms of Artifical Neural Network in that instead of focusing on the entirety of the problem domain, knowledge about the specific type of input is exploited. This in turn allows for a much simpler network architecture to be set up.

This paper has outlined the basic concepts of Convolutional Neural Networks, explaining the layers required to build one and detailing how best to structure the network in most image analysis tasks.

Research in the field of image analysis using neural networks has somewhat slowed in recent times. This is partly due to the incorrect belief surrounding the level of complexity and knowledge required to begin modelling these superbly powerful machine learning algorithms. The authors hope that this paper has in some way reduced this confusion, and made the field more accessible to beginners.

Training

CNNs and ANN in general use learning algorithms to adjust their free parameters in order to attain the desired network output. The most common algorithm used for this purpose is backpropagation. Backpropagation computes the gradient of an objective function to determine how to adjust a network's parameters in order to minimize errors that affect performance. A commonly experienced problem with training CNNs, and in particular DCNNs, is overfitting, which is poor performance on a held-out test set after the network is trained on a small or even large training set. This affects the model's ability to generalize on unseen data and is a major challenge for DCNNs that can be assuaged by regularization.

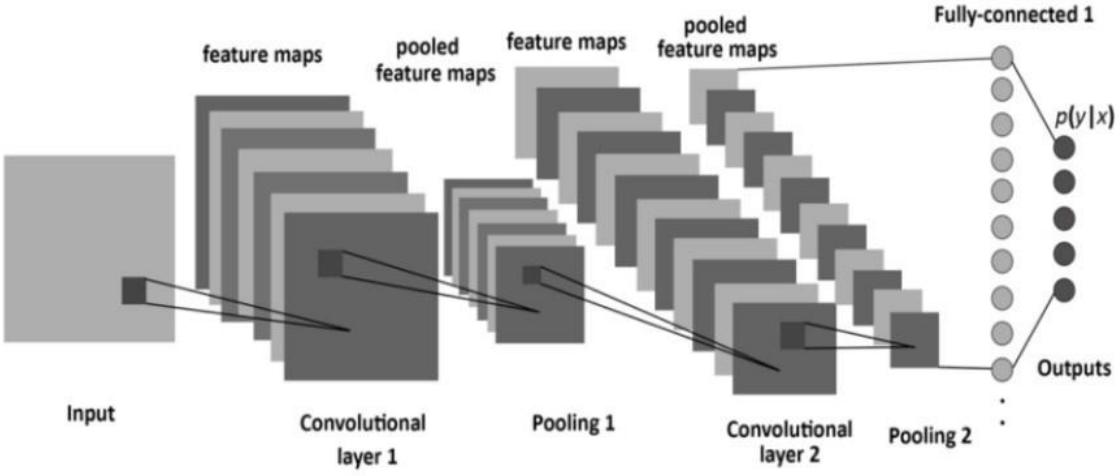


Fig: 4.5

Caffe Model

Caffe is a framework of Deep Learning and it was made used for the implementation and to access the following things in an object detection system.

•Expression: Models and optimizations are defined as plaintext schemas in the caffe model unlike others which use codes for this purpose.

•Speed: for research and industry alike speed is crucial for state-of-the-art models and massive data [11].

•Modularity: Flexibility and extension is majorly required for the new tasks and different settings.

•Openness: Common code, reference models, and reproducibility are the basic requirements of scientific and applied progress.

Types of Caffe Models

Open Pose

The first real-time multi-person system is portrayed by OpenPose which can collectively sight human body, hand, and facial keypoints (in total 130 keypoints) on single pictures.

Fully Convolutional Networks for Semantic Segmentation

In the absolutely convolutional networks (FCNs) Fully Convolutional Networks are the reference implementation of the models and code for the within the PAMI FCN and CVPR FCN papers.

Cnn-vis

Cnn-vis is an open-source tool that lets you use convolutional neural networks to generate images. It has taken inspiration from the Google's recent Inceptionism blog post.

Speech Recognition

Speech Recognition with the caffe deep learning framework.

DeconvNet

Learning Deconvolution Network for Semantic Segmentation.

Coupled Face Generation

This is the open source repository for the Coupled Generative Adversarial Network (CoupledGAN or CoGAN) work.These models are compatible with Caffe master, unlike earlier FCNs that required a pre-release branch (note: this reference edition of the models remains ongoing and not all of the models have yet been ported to master).

Codes for Fast Image Retrieval

To create the hash-like binary codes it provides effective framework for fast image retrieval.

SegNet and Bayesian SegNet

SegNet is real-time semantic segmentation architecture for scene understanding.

Deep Hand

It gives pre-trained CNN models.

DeepYeast

Deep Yeast may be an 11-layer convolutional neural network trained on biaural research pictures of yeast cells carrying fluorescent proteins with totally different subcellular localizations.

Python VS other languages for Object Detection: Object detection may be a domain-specific variation of the machine learning prediction drawback. Intel's OpenCV library that is implemented in C/C++ has its interfaces offered during a} very vary of programming environments like C#, Matlab, Octave, R, Python and then on. Why Python codes are much better option than other language codes for object detection are more compact and readable code.

Python uses zero-based indexing.

Dictionary (hashes) support provided.

Simple and elegant Object-oriented programming

Free and open

Multiple functions can be package in one module

More choices in graphics packages and toolsets Supervised learning also plays an important role.

The utility of unsupervised pre-training is usually evaluated on the premise of what performance is achieved when supervised fine-tuning. This paper reviews and discusses the fundamentals of learning

as well as supervised learning for classification models, and also talks about the mini batch stochastic gradient descent algorithm that is used to fine-tune many of the models.

Object Classification in Moving Object Detection Object classification works on the shape, motion, color and texture. The classification can be done under various categories like plants, objects, animals, humans etc. The key concept of object classification is tracking objects and analysing their features.

Shape-Based

A mixture of image-based and scene based object parameters such as image blob (binary large object) area, the as pectration of blob bounding box and camera zoom is given as input to this detection system. Classification is performed on the basis of the blob at each and every frame. The results are kept in the histogram.

Motion-Based

When an easy image is given as an input with no objects in motion, this classification isn't required. In general, non- rigid articulated human motion shows a periodic property; therefore this has been used as a powerful clue for classification of moving objects. based on this useful clue, human motion is distinguished from different objects motion. ColorBased- though color isn't an applicable live alone for police investigation and following objects, but the low process value of the colour primarily based algorithms makes the coloura awfully smart feature to be exploited. As an example, the color-histogram based technique is employed for detection of vehicles in period. Color bar chart describes the colour distribution in a very given region that is powerful against partial occlusions.

Texture-Based

The texture-based approaches with the assistance of texture pattern recognition work just like motion-based approaches. It provides higher accuracy, by exploitation overlapping native distinction social control however might need longer, which may be improved exploitation some quick techniques. I. proposed WORK Authors have applied period object detection exploitation deep learning and OpenCV to figure to work with video streams and video files. This will be accomplished using the highly efficient open computer vision. Implementation of proposed strategy includes caffe- model based on Google Image Scenery; Caffe offers the model definitions, optimization settings, pre- trained weights[4]. Prerequisite includes Python 3.7, OpenCV 4 packages and numpy to complete this task of object detection. NumPy is the elementary package for scientific computing with Python. It contains among other things: a strong N-dimensional array object, subtle (broadcasting) functions tools for integrating C/C++ and fortran code, helpful linear algebra, Fourier transform, and random

number capabilities. Numpy works in backend to provide statistical information of resemblance of object with the image scenery caffemodel database. Object clusters can be created according to fuzzy value provided by NumPy. This project can detect live objects from the videos and images.

LEARNING FEATURE HIERARCHY:

Learn hierarchy all the way from pixels classifier One layer extracts features from output of previous layer, train all layers jointly

Zero-One Loss

The models given in these deep learning tutorials are largely used for classification. The major aim of training a classifier is to reduce the amount of errors (zero-one loss) on unseen examples

Negative Log-Likelihood Loss

Optimizing it for large models (thousands or millions of parameters) is prohibitively expensive (computationally) because the zero-one loss isn't differentiable. In order to achieve this maximization of the log-likelihood is done on the classifier given all the labels in a training set [14].The likelihood of the correct class and number of right predictions is not the equal, but they are pretty similar from the point of view of a randomly initialized classifier. As the likelihood and zero-one loss are different objectives but we should always see that they are co-related on the validation set but sometimes one will rise while the other falls, or vice-versa.

Stochastic Gradient Descent

Ordinary gradient descent is an easy rule within which we repeatedly create tiny steps downward on an error surface defined by a loss function of some parameters. For the aim of normal gradient descent we take into account that the training data is rolled into the loss function. Then the pseudo code of this algorithm can be represented as Stochastic gradient descent (SGD) works according to similar principles as random gradient descent (SGD) operates on the basis of similar principles as normal gradient descent. It quickly proceeds by estimating the gradient from simply a few examples at a time instead of complete training set. In its purest kind, we use simply one example at a time to estimate the gradient.

Caffe is a deep learning framework or else we can say a library it's made with expression speed and modularity in mind they will put by Berkeley artificial intelligence research and created by young King Gia there are many deep learning or machine learning frameworks for computer vision like tensorflow ,Tiano, Charis and SVM[2]. But why exactly we implement edition cafe there as on is its expressive architecture we can easily switch between CPU and GPU while training on GPU machine

modules and optimization for Our problem is defined by configuration without hard coding. It supports extensible code since cafes are open source library. It is four foot by over twenty thous and developers and github since its birth it offers coding platform in extensible languages like Python and C++. The next reason is speed for training the neural networks speed is the primary constraint. Caffe can process over million images in a single day with the standard media GPU that is milliseconds per image. Whereas the same dataset of million images can take weeks for Tiana and Kara's Caffe is the fastest convolution neural network present community as mentioned earlier since its open source library huge number of research arepowered by cafe and every single day something new is coming out of it.

CHAPTER 5

OPEN COMPUTER VISION

5.1 INTRODUCTION

OpenCV stands for Open supply pc Vision Library is associate open supply pc vision and machine learning software system library. The purpose of creation of OpenCV was to produce a standard infrastructure for computer vision applications and to accelerate the utilization of machine perception within the business product [6]. It becomes very easy for businesses to utilize and modify the code with OpenCV as it is a BSD-licensed product. It is a rich wholesome libraby as it contains 2500 optimized algorithms, which also includes a comprehensive set of both classic and progressive computer vision and machine learning algorithms. These algorithms is used for various functions such as discover and acknowledging faces. Identify objects classify human actions. In videos, track camera movements, track moving objects. Extract 3D models of objects, manufacture 3D purpose clouds from stereo cameras, sew pictures along to provide a high-resolution image of a complete scene, find similar pictures from a picture information, remove red eyes from images that are clicked with the flash, follow eye movements, recognize scenery and establish markers to overlay it with augmented reality.

Officially launched in 1999 the OpenCV project was initially an Intel Research initiative to advance CPU-intensive applications, part of a series of projects including real-time ray tracing and 3D display walls The main contributors to the project included a number of optimization experts in Intel Russia, as well as Intel's Performance Library Team. In the early days of OpenCV, the goals of the project were describedas:

- Advance vision research by providing not only open but also optimized code for basic vision infrastructure. No more reinventing the wheel.
- Disseminate vision knowledge by providing a common infrastructure that developers could build on, so that code would be more readily readable and transferable.
- Advance vision-based commercial applications by making portable, performance-optimized code available for free – with a license that did not require code to be open or free itself.

The first alpha version of OpenCV was released to the public at the IEEE Conference on Computer Vision and Pattern Recognition in 2000, and five betas were released between 2001 and 2005. The first 1.0 version was released in 2006. A version 1.1 "pre-release" was released in October 2008.

The second major release of the OpenCV was in October 2009. OpenCV 2 includes major changes to the C++ interface, aiming at easier, more type-safe patterns, new functions, and better implementations for existing ones in terms of performance (especially on multi-core systems). Official releases now occur every six months and development is now done by an independent Russian team supported by commercial corporations.

In August 2012, support for OpenCV was taken over by a non-profit foundation OpenCV.org, which maintains a developer and user site.

On May 2016, Intel signed an agreement to acquire ITSEEZ, a leading developer of OpenCV.

OpenCV (Open source computer vision) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source BSD license.

It has C++, Python, Java and MATLAB interfaces and supports Windows, Linux, Android and Mac OS. OpenCV leans mostly towards real-time vision applications and takes advantage of MMX and SSE instructions when available. A full-featured CUDAandOpenCL interfaces are being actively developed right now.

There are over 500 algorithms and about 10 times as many functions that compose or support those algorithms. OpenCV is written natively in C++ and has a templated interface that works seamlessly with STL containers.

OpenCV's application areas include :

♣ 2D and 3D feature toolkits

♣Egomotion estimation

♣ Facial recognition system

♣ Gesture recognition

♣ Human–computer interaction (HCI)

♣ Mobile robotics

♣ Motion understanding

♣ Object identification

♣ Segmentation and recognition

♣ Stereopsis stereo vision: depth perception from 2 cameras

♣ Structure from motion (SFM)

♣ Motion tracking

♣ Augmented reality

To support some of the above areas, OpenCV includes a statistical machine learning library that contains :

♣ Boosting Decision tree learning

♣ Gradient boosting trees

♣ Expectation-maximization algorithm

♣ k-nearestneighbor algorithm

♣ Naive Bayes classifier

♣ Artificial neural networks

♣ Random forest

♣ Random forest

♣ Support vector machine (SVM)

♣ Deep neural networks (DNN)

Libraries in OpenCV

Numpy:

NumPy is an acronym for "Numeric Python" or "Numerical Python". It is an open source extension module for Python, which provides fast precompiled functions for mathematical and numerical routines. Furthermore, NumPy enriches the programming language Python with powerful data structures for efficient computation of multi-dimensional arrays and matrices. The implementation is even aiming at huge matrices and arrays. Besides that the module supplies a large library of high-level mathematical functions to operate on these matrices and arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

♣ A powerful N-dimensional array object Sophisticated (broadcasting) functions

♣ Tools for integrating C/C++ and Fortran code

♣Useful linear algebra, Fourier Transform, and random number capabilities.

Numpy Array:

A numpy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

SciPy:

SciPy (Scientific Python) is often mentioned in the same breath with NumPy. SciPy extends the capabilities of NumPy with further useful functions for minimization, regression, Fourier-transformation and many others. NumPy is based on two earlier Python modules dealing with arrays. One of these is Numeric. Numeric is like NumPy a Python module for high-performance, numeric computing, but it is obsolete nowadays. Another predecessor of NumPy is Numarray, which is a complete rewrite of Numeric but is deprecated as well. NumPy is a merger of those two, i.e. it is build on the code of Numeric and the features of Numarray.

The Python Alternative To Matlab :

Python in combination with Numpy, Scipy and Matplotlib can be used as a replacement for MATLAB. The combination of NumPy, SciPy and Matplotlib is a free (meaning both "free" as in "free beer" and "free" as in "freedom") alternative to MATLAB. Even though MATLAB has a huge number of additional toolboxes available, NumPy has the advantage that Python is a more modern and complete programming language and - as we have said already before - is open source. SciPy adds even more MATLAB-like functionalities to Python. Python is rounded out in the direction of MATLAB with the module Matplotlib, which provides MATLAB-like plotting functionality.

Haar Cascade Classifier in OpenCv

The algorithm needs a lot of positive images (images of faces) and negative images (images without faces) to train the classifier. Then we need to extract features from it. For this, haar features shown in below image are used. They are just like our convolutional kernel. Each feature is a single value obtained by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle.

Now all possible sizes and locations of each kernel is used to calculate plenty of features. (Just imagine how much computation it needs? Even a 24x24 window results over 160000 features). For each feature calculation, we need to find sum of pixels under white and black rectangles. To solve this, they

introduced the integral images. It simplifies calculation of sum of pixels, how large may be the number of pixels, to an operation involving just four pixels. Nice, isn‟t it? It makes things super-fast.

But among all these features we calculated, most of them are irrelevant. For example, consider the image below. Top row shows two good features. The first feature selected seems to focus on the property that the region of the eyes is often darker than the region of the nose and cheeks. The second feature selected relies on the property that the eyes are darker than the bridge of the nose. But the same windows applying on cheeks or any other place is irrelevant. So how do we select the best features out of 160000+ features? It is achieved by Adaboost.

For this, we apply each and every feature on all the training images. For each feature, it finds the best threshold which will classify the faces to positive and negative. But obviously, there will be errors or misclassifications. We select the features with minimum error rate, which means they are the features that best classifies the face and non-face images. (The process is not as simple as this. Each image is given an equal weight in the beginning. After each classification, weights of misclassified images are increased. Then again same process is done. New error rates are calculated. Also new weights. The process is continued until required accuracy or error rate is achieved or required number of features are found).

Final classifier is a weighted sum of these weak classifiers. It is called weak because it alone can‟t classify the image, but together with others forms a strong classifier. The paper says even 200 features provide detection with 95% accuracy. Their final setup had around 6000 features. (Imagine a reduction from 160000+ features to 6000 features. That is a big gain).

So now you take an image. Take each 24x24 window. Apply 6000 features to it. Check if it is face or not. Wow.. Wow..Isn‟t it a little inefficient and time consuming? Yes, it is. Authors have a good solution for that.

In an image, most of the image region is non-face region. So it is a better idea to have a simple method to check if a window is not a face region. If it is not, discard it in a single shot. Don‟t process it again. Instead focus on region where there can be a face. This way, we can find more time to check a possible face region

For this they introduced the concept of Cascade of Classifiers. Instead of applying all the 6000 features on a window, group the features into different stages of classifiers and apply one-by-one. (Normally first few stages will contain very less number of features). If a window fails the first stage, discard it. We don‟t consider remaining features on it. If it passes, apply the second stage of features and

continue the process. The window which passes all stages is a face region. Haar-like features are digital image features used in object recognition. They owe their name to their intuitive similarity with Haar wavelets and were used in the first real-time face detector.

Historically, working with only image intensities (i.e., the RGB pixel values at each and every pixel of image) made the task of feature calculation computationally expensive. A publication by Papageorgiou et al. discussed working with an alternate feature set based on Haar wavelets instead of the usual image intensities. Paul Viola and Michael Jones adapted the idea of using Haar wavelets and developed the so-called Haar-like features. A Haar-like feature considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. For example, with a human face, it is a common observation that among all faces the region of the eyes is darker than the region of the cheeks. Therefore, a common Haar feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region. The position of these rectangles is defined relative to a detection window that acts like a bounding box to the target object (the face in this case).

In the detection phase of the Viola–Jones object detection framework, a window of the target size is moved over the input image, and for each subsection of the image the Haar-like feature is calculated. This difference is then compared to a learned threshold that separates non-objects from objects. Because such a Haar-like feature is only a weak learner or classifier (its detection quality is slightly better than random guessing) a large number of Haar-like features are necessary to describe an object with sufficient accuracy. In the Viola–Jones object detection framework, the Haar-like features are therefore organized in something called a classifier cascade to form a strong learner or classifier.

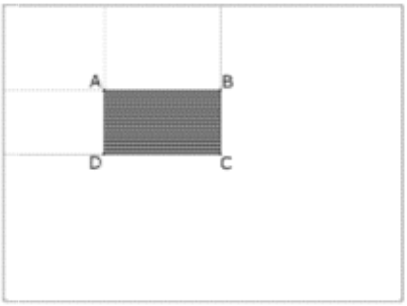
The key advantage of a Haar-like feature over most other features is its calculation speed. Due to the use of integral images, a Haar-like feature of any size can be calculated in constant time (approximately 60 microprocessor instructions for a 2-rectangle feature).

Rectangular Haar-like features

A simple rectangular Haar-like feature can be defined as the difference of the sum of pixels of areas inside the rectangle, which can be at any position and scale within the original image. This modified feature set is called 2-rectangle feature. Viola and Jones also defined 3-rectangle features and 4-rectangle features. The values indicate certain characteristics of a particular area of the image. Each feature type can indicate the existence (or absence) of certain characteristics in the image, such as edges or changes in texture. For example, a 2-rectangle feature can indicate where the border lies between a dark region and a light region.

Fast Computation of Haar-like features:

One of the contributions of Viola and Jones was to use summed-area tables, which they called integral images. Integral images can be defined as two-dimensional lookup tables in the form of a matrix with the same size of the original image. Each element of the integral image contains the sum of all pixels located on the up-left region of the original image (in relation to the element's position).



Sum= I(C) + I(A) – I(B) – I(D)

Fig: 5.1 2-Rectangle feature

OpenCV has a modular structure, which means that the package includes several shared or static libraries. The following modules are available:

- Core functionality (core) - a compact module defining basic data structures, including the dense multi-dimensional array Mat and basic functions used by all other modules.
- Image Processing (imgproc) - an image processing module that includes linear and non-linear image filtering, geometrical image transformations (resize, affine and perspective warping, generic table-based remapping), color space conversion, histograms, and so on.
- Video Analysis (video) - a video analysis module that includes motion estimation, background subtraction, and object tracking algorithms.
- Camera Calibration and 3D Reconstruction (calib3d) - basic multiple-view geometry algorithms, single and stereo camera calibration, object pose estimation, stereo correspondence algorithms, and elements of 3D reconstruction.
- 2D Features Framework (features2d) - salient feature detectors, descriptors, and descriptor matchers.
- Object Detection (objdetect) - detection of objects and instances of the predefined classes (for example, faces, eyes, mugs, people, cars, and so on).
- High-level GUI (highgui) - an easy-to-use interface to simple UI capabilities.

- Video I/O (videoio) - an easy-to-use interface to video capturing and video codecs.
- Some other helper modules, such as FLANN and Google test wrappers, Python bindings, and others.

CHAPTER 6

RESULTS AND DISCUSSIONS

INTRODUCTION TO IMPLEMENTATION OF PROBLEM

The Model

Deep learning is a popular technique used in computer vision. We chose Convolutional Neural Network (CNN) layers as building blocks to create our model architecture. CNNs are known to imitate how the human brain works when analyzing visuals.

A typical architecture of a convolutional neural network contain an input layer, some convolutional layers, some dense layers (aka. fully-connected layers), and an output layer . These are linearly stacked layers ordered in sequence.

Input Layer

The input layer has pre-determined, fixed dimensions, so the image must be pre-processed before it can be fed into the layer. We used OpenCV, a computer vision library, for object detection in the video.

The OpenCV contains pre-trained filters and uses Adaboost to quickly find and crop the object. The cropped object is then converted into gray scale using cv2.cvtColor and resized to 48-by-48 pixels with cv2.resize. This step greatly reduces the dimensions compared to the original RGB format with three colour dimensions (3, 48, 48). The pipeline ensures every image can be fed into the input layer as a (1, 48, 48) numpy array.

Convolutional Layers

The numpy array gets passed into the Convolution2D layer where we specify the number of filters as one of the hyper parameters. The set of filters are unique with randomly generated weights. Each filter, (3, 3) receptive field, slides across the original image with shared weights to create a feature map.

Convolution generates feature maps that represent how pixel values are enhanced, for example, edge and pattern detection. A feature map is created by applying filter 1 across the entire image. Other filters are applied one after another creating a set of feature maps.

Pooling is a dimension reduction technique usually applied after one or several convolutional layers. It is an important step when building CNNs as adding more convolutional layers can greatly affect computational time. We used a popular pooling method called MaxPooling2D that uses (2, 2) windows across the feature map only keeping the maximum pixel value. The pooled pixels form an image with dimentions reduced by 4.

Dense Layers

The dense layer (aka fully connected layers), is inspired by the way neurons transmit signals through the brain. It takes a large number of input features and transform features through layers connected with trainable weights.

These weights are trained by forward propagation of training data then backward propagation of its errors. Back propagation starts from evaluating the difference between prediction and true value, and back calculates the weight adjustment needed to every layer before. We can control the training speed and the complexity of the architecture by tuning the hyper-parameters, such as learning rate and network density. As we feed in more data, the network is able to gradually make adjustments until errors are minimized.Essentially, the more layers/nodes we add to the network the better it can pick up signals.

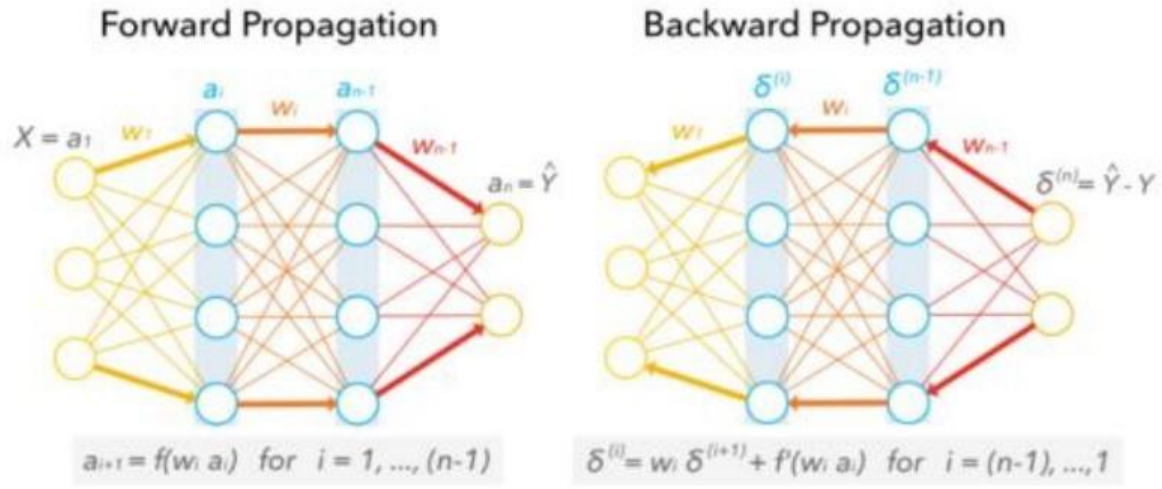


Fig: 6.1

As good as it may sound, the model also becomes increasingly prone to overfitting the training data. One method to prevent overfitting and generalize on unseen data is to apply dropout. Dropout randomly selects a portion (usually less than 50%) of nodes to set their weights to zero during training. This method can effectively control the model's sensitivity to noise during training while maintaining the necessary complexity of the architecture.

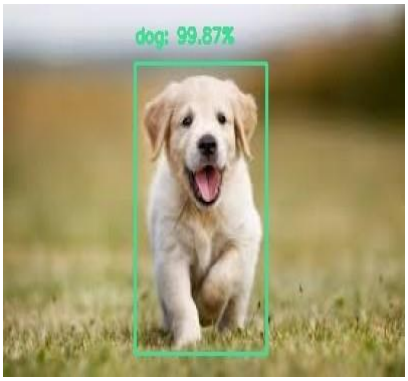
Output layer

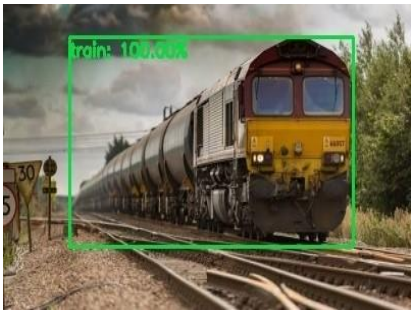
The output layer in a CNN as mentioned previously is a fully connected layer, where the input from the other layers is flattened and sent so as the transform the output into the number of classes as desired by the network.

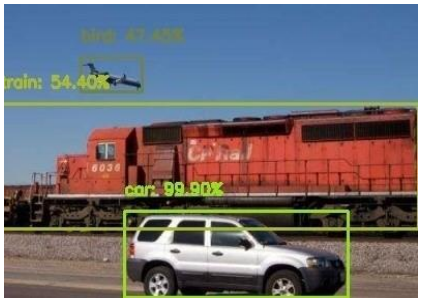
RESULTS

Input

Output







CONCLUSION:

Deep learning based object detection has been a research hotspot in recent years. This project starts on generic object detection pipelines which provide base architectures for other related tasks. With the help of this the three other common tasks, namely object detection, face detection and pedestrian detection, can be accomplished. Authors accomplished this by combing two things: Object detection with deep learning and OpenCV and Efficient, threaded video streams with OpenCV. The camera sensor noise and lightening condition can change the result as it can create problem in recognizing the object. The end result is a deep learning- based object detector that can process around 6-8 FPS.

REFERENCES:

Bruckner, Daniel. Ml-o-scope: a diagnostic visualization system for deep machine learning pipelines. No. UCB/EECS-2014-99.CALIFORNIA UNIV BERKELEY DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES, 2014.

K Saleh, Imad, Mehdi Ammi, and Samuel Szoniecky, eds. Challenges of the Internet of Things: Technique, Use, Ethics. John Wiley & Sons, 2018.

Petrov, Yordan. Improving object detection by exploiting semantic relations between objects.MS thesis.UniversitatPolitècnica de Catalunya, 2017.

Nikouei, SeyedYahya, et al. "Intelligent Surveillance as an Edge Network Service: from Harr-Cascade, SVM to a Lightweight CNN." arXiv preprint arXiv:1805.00331 (2018).

Thakar, Kartikey, et al. "Implementation and analysis of template matching for image registration on DevKit- 8500D." Optik-International Journal for Light and Electron Optics 130 (2017): 935-944..

Bradski, Gary, and Adrian Kaehler.Learning OpenCV: Computer vision with the OpenCV library." O'Reilly Media, Inc.", 2008.

Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

Kong, Tao, et al. "Ron: Reverse connection with objectness prior networks for object detection." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2017.

Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision.Springer, Cham, 2016.

Veiga, Francisco José Lopes. "Image Processing for Detection of Vehicles In Motion." (2018).

Huaizheng Zhang, Han Hu, GuanyuGao, Yonggang Wen, Kyle Guan, "Deepqoe: A Unified Framework for Learning to Predict Video QoE", Multimedia and Expo (ICME) 2018 IEEE International Conference on, pp. 1- 6, 2018.

Shijian Tang and Ye Yuan,“Object Detection based on Conventional Neural Network”.

R. P. S. Manikandan, A. M. Kalpana, "A study on feature selection in big data", Computer Communication and Informatics (ICCCI) 2017 International Conference on, pp. 1-5, 2017

Warde-Farley, David. "Feedforward deep architectures for classification and synthesis." (2018).

Shilpisingh et al" An Analytic approach for 3D Shape descriptor for face recognition", International Journal of Electrical, Electronics, Computer Science & Engineering (IJEECSE), Special Issue - ICSCAAIT-2018| E-ISSN: 2348-2273 | P-ISSN: 2454-1222,pp-138-140.

REAL TIME OBJECT DETECTION WITH DEEP LEARNING

Ch. Padma Sree[1], K. Sai Naveen[2], J.A.S. Sampreeth[3], D. Pavan[4], V. S. Ashlesh Kumar[5]
Department of Electronics and Communications Engineering, Andhra University, ANITS College, Visakhapatnam, India
padmasreechalla8@gmail.com
kandarpanaveen7@gmail.com
pavanvidyasagar999@gmail.com

Abstract— Real time object detection is a vast, vibrant and sophisticated area of computer vision aimed towards object identification and recognition. Object detection detects the semantic objects of a class objects using OpenCV (Open source Computer Vision), which is a library of programming functions mainly trained towards real time computer vision in digital images and videos. Visually challenged people cannot distinguish the objects around them. The main aim behind this real time object detection is to help the blind to overcome their difficulty. Real time object detection finds its uses in the areas like tracking objects, video surveillance, pedestrian detection, people counting, self-driving cars, face detection, ball tracking in sports and many more. This is achieved using Convolution Neural Networks, which is a representative tool of Deep learning. This project acts as an aiding tool for visually challenged people.

Keywords: Convolutional Neural Network, OpenCV, Deep Learning.

I. INTRODUCTION

Object detection is a technology to detect various objects in digital images and videos too. It is mainly helpful within the self- driving cars, face detection, etc., where the objects are to be continuously monitored. The algorithm or the technique involved for object detection during this project is Convolutional Neural Networks which is a class of Deep learning. This uses MobileNet SSD technique during which MobileNet is a neural network used for image classification and recognition whereas SSD is a framework that is used to realize the multibox detector. The mixture of both MobileNet and SSD can do object detection. The main advantage or purpose of choosing Deep learning is that we do not need to do feature extraction from data as compared to machine learning.

The Haar-like trait play a crucial role in detecting the objects in a picture. They scan the entire picture starting from the top left and compares every small box with the trained data. In this way, even small-detailed objects present within the images are identified.

I **II. METHODOLOGY**

Deep learning, a subset of machine learning which in turn is a subset of artificial intelligence (AI) has networks capable of learning things from the data that is unstructured or unlabeled. The approach utilized in this project is Convolutional Neural Networks (CNN). It uses the Haar-cascade classifiers which help us in the detection of objects.

1. CNN:

The convolutional neural network, or CNN for brief, could also be a specialized kind of neural network model designed for working with two-dimensional image data, although they're going to be used with one-dimensional and three-dimensional data.

Central the convolutional neural network is the convolutional layer that gives the network its name. This layer performs an operation known as “convolution”.

In the context of a convolutional neural network, a convolution may be a linear operation that involves the multiplication of a group of weights with the input, very similar to a standard neural network. as long as the technique was designed for two-

dimensional input, the multiplication is performed between an array of input file and a two-dimensional array of weights, called a filter or a kernel.

The filter is smaller than the input file and therefore the before the sort of multiplication applied between a filter-sized patch of the input and the filter may be a scalar product. A scalar product is

that the element-wise multiplication between the filter-sized patch of the input and filter, which is then summed, always leading to one value. Because it leads to 1 value, the operation is conventionally represented and mentioned because the "scalar product".

Using a filter smaller than the input is intentional because it allows an equivalent filter (set of weights) to be multiplied by the input array multiple times at distinct points on the input. Specifically, the filter is applied systematically to every overlapping part or filter-sized patch of the input file, left to right, top to bottom.

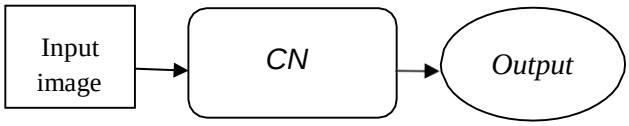


Fig 1.1 Sample block diagram indicating the flow of image processing using CNN

This systematic application of an equivalent filter across a picture may be a powerful idea. If the filter is meant to detect a selected sort of feature within the input, then the appliance of that filter systematically across the whole input image allows the filter a chance to get that feature anywhere within the image.

This capability is usually represented and mentioned as translation invariance, e.g. the total altogether concern in whether the feature is present instead of where it should had been present.

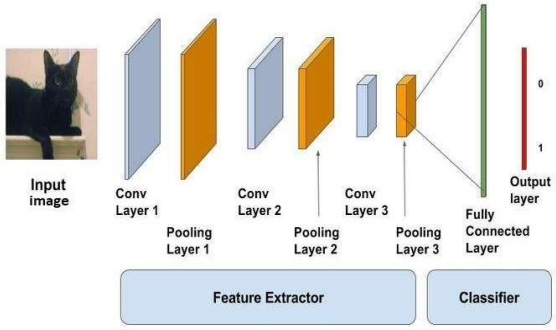


Fig1.2. Image classification using CNN

2. OpenCV:

Open CV stands for open source computer vision. it's a group of libraries in Python. it's a tool by which we will be able to manipulate the pictures , like image scaling, etc. This supports and helps us in developing real time computing applications. It mainly concentrates and targets on image processing, video capture and analysis. It includes several features like face detection and also object detection. Currently OpenCV supports differing types of programming languages like C++, Python, Java etc., and it's available on various platforms including Windows, Linux, OS X, Android etc.

3. Training the data set:

The data set is typically the gathering of knowledge . the info set could also be collection of images or alphabets or numbers or documents and files too. the info set we used for the thing detection is that the collection of images of all the objects that are to be identified. Several different images of every and each object is typically present within the data set. If there are more number of images like each object within the datasets then the accuracy are often improved. The important thing that's to be remembered is that the info within the data set must be labelled. there'll be actually 3 data set. they're the training data set, the validation dataset and therefore the other one is testing data set. The training data set will usually contains around 85-90% of the entire labelled data. This training dataset are going to be training our machine and therefore the model is obtained by training the info set. The validation data set consists of around 5-10% of the entire labelled data. this is often used for the validation purpose. the opposite data set is that the testing dataset and it's wont to test the performance of our machine.

4. Developing a real time object detector:

For developing a true time object detector using deep learning and open cv we'd like to access our web cam during a really effective way then the thing detection is to be applied to each and every frame. we should always install open cv in our systems.

The deep neural network module should be installed.

Firstly, we should always always import all the specified packages:

1. From imutils.video we'll import VideoStream

2. From imutils.video we'll import FPS

3. we'll import numpy as np

4. we'll import argparse

5. we'll import imutils

6. we'll import time

7. we'll import cv2

The next step is to construct the argument parse then we should always parse the arguments.

--prototxt: provide path to the Caffe prototxt file.

--model: provide path to the pre-trained model.

--confidence: The minimum probability threshold to filter weak detections. The default value is given as 20%.

The next step is to initialize CLASS labels and corresponding random COLORS.

Each object when it's detected, it's surrounded by a box with some predefined colour. Thus, we assign each object a specific color.

After that we'll load our model and that we will provide the regard to our prototxt and also to our model files.

With the assistance of imutils we'll read the video and that we will set the amount of frames per second. Now with this some predefined number of frames are going to be loaded per second. Each frame is analogous to the image. Now these images are going to be given because the inputs to the model.

The model will process the input image and produces the output image which consists of labels. in additional practical sense the input raw image is given to the model. Now the model process the input image. within the output image all the

thing s are identified and every object is surrounded by an oblong box and therefore the name of the object is additionally displayed. we'll be only observing the output video stream but not the input video stream.

III. RESULT

Here, in this project we've considered around 15 to 20 objects to be detected during the training. Some of those include 'person', 'car', 'train', 'bird', 'sofa', 'dog', ''plant', 'aero plane', 'bicycle', 'bus', 'motorbike', etc.

The output of this project displays the objects detected with a rectangular box around the object with a label indicating it's name and therefore the exactness with which the object has been detected on the top of it. It can dig out any number of objects existing during a single image with certainty.

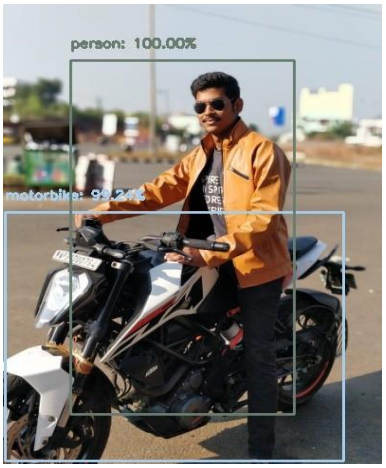


Fig 1

Fig 2

Fig 3

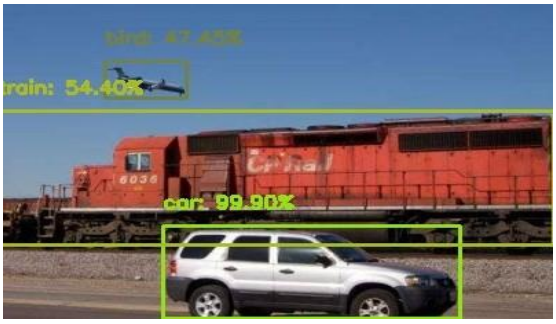


Fig 4

Fig 5

IV. APPLICATIONS

Here are a some of the future implementation of object detection.

1. Face detections and recognition:

Face detection perhaps be a separate class of object detection. We wonder how some applications like Facebook, Faceapp, etc., detect and recognize our faces. this is often a sample example of object detection in our day to day life. Face detection is already in use in our lifestyle to unlock our mobile phones and for other security systems to scale back rate .

2. Object tracking:

Object detection is additionally utilized in tracking objects like tracking an individual and his actions, continuously monitoring a ball within the game of Football or Cricket. As there's an enormous interest for people in these games, these tracking techniques enables them to know it during a better way and obtain some additional information. Tracking of the ball is of maximal importance in any ball-based games to automatically record the movement of the ball and adjust the video frame accordingly.

3. Self-driving cars:

this is often one among the main evolutions of the planet and is that the best example why we'd like object detection. so as for a car to travel to the specified destination automatically with none human interference or to form decisions whether to accelerate or to use brakes and to spot the objects around it. this needs object detection.

4. Emotions detection:

this permits the system to spot the type of emotion the person puts on his face. the corporate Apple has already tried to use this by detecting the emotion of the user and converting it into a respective emoji within the smart phone.

5. Biometric identification through retina scan:

Retina scan through iris code is one among the techniques utilized in high security systems because it is one among the foremost accurate and unique biometric.

6. Smart text search and text selection (Google lens)

In recent times, we've encountered an application in smart phones called google lens. this will recognize the text and also images and search the relevant information within the browser without much effort.

V. CONCLUSION

Deep-learning based object detection has been a search hotspot in recent years. This project starts on generic object detection pipelines which give base architectures for other related tasks. With the assistance of this the 3 other common tasks, namely object detection, face detection and pedestrian detection, are often accomplished. Authors accomplished this by combing 2 things: Object detection with deep learning and OpenCV and Efficient, threaded video streams with OpenCV. The camera sensor noise and lightening condition can change the result because it can create problem in recognizing the objects. generally, this whole process requires GPU's rather than CPU's. But we've done using CPU's and executes in much less time, making it efficient. Object Detection algorithms act as a mixture of both image classification and object localization. It takes the given image as input and produces the output having the bounding boxes adequate to the amount of objects present within the image with the category label attached to every bounding box at the highest. It projects the scenario of the bounding box up the shape of position, height and width.

VI. REFERENCES

1. Geethapriya S, N. Duraimurugan, S.P. Chokkalingam, "Real-Time Object Detection with Yolo", International Journal of Engineering and Advanced Technology (IJEAT)
2. Abdul Vahab, Maruti S Naik, Prasanna G Raikar an Prasad S R4, "Applications of Object Detection System", International Research Journal of Engineering and Technology (IRJET)
3. Hammad Naeem, Jawad Ahmad and Muhammad Tayyab, "Real-Time Object Detection and Tracking",IEEE
4. Meera M K, & Shajee Mohan B S. 2016, "Object recognition in images", International Conference on Information

Science (ICIS).
5. Astha Gautam, Anjana Kumari, Pankaj Singh: "The Concept of Object Recognition", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015
6. Joseph Redmon, Santosh Divvala, Ross Girshick, “You Only Look Once: Unified, Real-Time Object Detection”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016,pp. 779-788
7. V. Gajjar, A. Gurnani and Y. Khandhediya, "Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach," in 2017 IEEE International Conference on Computer Vision Workshops, 2017.