ABSTRACT

In collaboration with two vast fields of digital world i.e., Machine Learning and Artificial Intelligence; the further contents describe about the project based on the topic "**To Identify and Recognize the Object for Traffic Analysis System using Deep Learning".** In addition to Object recognition; implementation of this technique under traffic analysis system was also under consideration. This project has brought human work to an end, accurate clarity to object classification, to get a better view on the classification of objects, and overall of this advantage - it has added a huge step to this new digital world. The objective of this project is to enhance the accuracy of object identification and reduce the evaluation time in different environment. Above all this to give a clarity to traffic control system and its advancement . The methods and techniques used for this project is **"YOLO"** algorithm and "**Convolutionary Neural Networks".** The entire project is classified into different modules i.e., Multiple grid system , Square dimension system , hardware requirements "Web Cam for recording , Operating System with appropriate specification" , and software requirements "right libraries (pandas, matplotlib, TensorFlow, Numpy, OpenCV-python, Sklearn , imgaug) and frameworks". The project leads to less execution time in GPU and CPU environment . Public security and advancement in traffic system is a proof to its significance .

INTRODUCTION

1.1. Overview

Over recent years, Computer Vision along with advanced version of deep learning enhanced the role of human resources for overall development. Computer vision is the method to extract the details of a digital image which can be easily done through our eyes. Image processing with the help of Convolution Neural Networks provided a new name to digital world of enhancement. In addition, it supports the applications of computer vision such as Object detection, Face identification, Object tracking, Semantic segmentation. Object recognition is evolving from the single object detection to the multi-object detection. Object detection accompanied by Artificial Intelligence can be used in smart auto-vehicles. The proposed system can be implemented in the vehicles which used for object-detection. The system is so designed to identify road signs and pedestrian on the roads and the system algorithm will work on it with prepared dataset. It will identify the signs and passes the signals to appropriate vehicle machinery or to the driver for further respond to that particular incident. The Object Identification operation is based on CNN – algorithm and these popular datasets are upgrading to the highest accuracy. In this paper, we first describe some algorithms associated with deep learning to identify objects, then apply an algorithm to a new dataset to check its availability. Deep learning permits models are made of multiple embedding layers for learning abstraction from many levels of abstraction. These modes improve speech recognition, visual object recognition, object oriented and many other domains such as modern state Drug discovery and genomics. In-depth training explores the complex structure of large data sets using a back-propagation algorithm to demonstrate how a machine must alter its internal parameters to represent each layer in the back-to-back representation. Go. The deep Convolunary Network improves the processing of images, video, speech and audio, while continuous networks shed light on the gains of data such as text and speech. For starters it may be challenging to differentiate between different computerized computer vision tasks. For example, the image classification is quite obvious, but the difference between the object's localization and the object's identity can be confusing, especially when all three functions can be changed in the same way as object recognition. Image classification affects assigning an image to a class label, while object localization involves drawing a bounding box around one or more objects in an image. Identifying objects is more challenging and combines both functions and assigns a class label to a bounding box around each object of interest in the image. Together, all these

challenges are related to unit recognition. The first intention of trying to "understand the scene" is one of the base ideas in computer vision that lead to a continuous increase in the need to apprehend the high-level context in images regarding object recognition and image classification. By becoming a fundamental visual expertise that Computer Vision systems require, the field has rapidly grown. Images have become ubiquitous in a variety of fields as so many people and systems extract vast amounts of information from imagery. Information that can be vital in areas such as robotics, hospitals, self-driving cars, surveillance or building 3D representations of objects. While each of the above-mentioned applications differs by numerous factors, they share the common process of correctly annotating an image with one or a probability of labels that correlates to a series of classes or categories. This procedure is known as image classification and, combined with machine learning, it has become an important research topic in the field, on account of the focus on the understanding of what an image is representative of. The complex process of identifying the type of materials in diverse tasks linked to image-based scene perspectives has taken advantage of the combination of machine learning techniques applied to the up-to-date development of neural networks. This outlines the challenging problem of material classification due to the variety of the definite features of materials. The state of-the-art solutions rely massively on the attention that Computer Vision systems have received, which led to a series of algorithms being developed and images being collected in datasets.

People are able to recognize the environment they are in as well as the various objects in their everyday life no matter the influence on the item's features or if their view is obstructed, as this is one of the very first skills, we learn from the moment we are born. Computers, on the other hand, require effort and powerful computation and complex algorithms to attempt to recognize correctly patterns and regions where a possible object might be. Object detection and recognition are two main ways that have been implemented over multiple decades that are at the center of Computer Vision systems at the moment. These approaches are presented with challenges such as scale, occlusion, view point, illumination or background clutter, all issues that have been attempted as research topics that provided functionality that led to the introduction of Neural Networks and Convolutional Neural Networks (CNN). The newly added functionality is composed of distinct types of layers that consist of many parameters that are able to figure out the features present in a given image. These architectures have since been built on and a more complex structure with hidden non-linear layers between the input and output layers of a CNN has been identified as Deep Convolutional Neural Network (DCNN).

The advancement in the computational speed of computers and the amount of data available has allowed deep learning to increase the overall performance under supervised learning conditions. Results have been shown to be better when there is more data available used by bigger models on a faster system. favouring the latest research, scientists brought into attention two new topics: object localization and semantic segmentation.

The first step of approaching the demanding issue of image classification is by looking at the available data. Material datasets started to be gathered just over a decade ago as only a few numbers of computer vision systems targeted material recognition as a research topic. Analysing Columbia Utrecht Reflectance and Texture Database CUReT [8], the first published dataset, it can be observed that the environment conditions are limited. Although it attempts its best at simulating the real-world settings through having under 205 viewings and an extensive amount illumination directions and having a performance greater than 95%, the fact that it uses synthetic data and a very low number of images per category makes the results inaccurate and limited in capturing the complexity of the materials found in the real-world [9]. This conclusion was emphasized in a low accuracy of just 23% obtained when the same approach was used on the Flickr Materials Dataset (FMD) produced by Sharan et al [10]. The main noticeable difference was the fact that FMD database was using real-world images and increased the number of images per category to 100 compared to CUReT dataset that only had 61 images. The Open Surfaces dataset is formed of real-world images and achieves a 34.5% accuracy. This dataset takes a step further and focuses on objects from consumer photographs and introduces a larger database with 53 classes and 25000 scenes with more than 110000 segmented materials. An important achievement in the field was presented shortly by Liu et al when the issue of recognizing the type of material from its own features was tackled and a better accuracy of 45% was found. This offered the start for extended research into the areas of shape-based object recognition or texture recognition by developing an algorithm that successfully finds the object and extracts features like its shape, colour, texture and reflectance. The algorithm uses a technique called Bag-of-Visual-Words (BoVW) where the words are defined by the features extracted and the bag-of-words is represented by the picture. Cimpoi et al have later released the Describable Textures Dataset (DTD) where for the first time an Improved Fisher Vector (IFV) and a CNN architecture have been combined to outline that together the two approaches had better found the key characteristics of objects by outperforming previous work by 10% and becoming the state-of-the-art. With deep learning achieving better results, Gibert et al proved that by using DCNNs specifically trained for the

task of material recognition on railway tracks, the system produces an output of 93.35% efficiency and that such systems continue to produce the state-of-the-art results for image classification and object detection. In the same year, Bell et al, has achieved results of 85.2% mean accuracy using CNNs and a brand-new dataset called Materials in Context Database (MINC) that has around 2500 images per category and has a total of 23 classes.

In Computer Vision systems, datasets are divided into two main categories: a training dataset used for training the algorithm learn to perform its desired task and a testing dataset that the algorithm is tested on. The percentages that one divides them by affects the general pipeline and it is the first step when trying to solve this challenging task. The work that is being offered within this report, focuses on a 50% split across all datasets tested. Furthermore, each training and testing dataset is further split into negative and positive structures in order to make sure that the algorithm learns what it should be looking for in an image as well as what not to look for. At the end of a test run, one needs to look at all accuracy counts to be able to come up with a total analysis of the pipeline created. This need emphasizes the fact that the system learns how to better optimize itself. Using previous research work as a base, a comparison of several types of CNN architectures is made. To have as conclusive results a complex and full analysis as well as making sure that the learning process is beneficial for the entire pipeline, and particularly for the feature extractors, four widely used main datasets have been selected for training and testing purposes.

Transfer learning is a promising technique that improves the model learning and is used for object recognition when there are very few training samples and when the need for analyse data needs to be reduced. This approach can be compared to being an augmentation as the information is transferred between layers or classes. In the work by Lim et al, it is used for object detection where the developed framework borrows examples from categories in order to learn how to transform those examples to become more similar to the instances from the targeted class. Taking advantage of the information already available in other classes leads to better results for object recognition in which try to adapt the classifiers from the pipeline to overcome the fact that there are not enough training images. By receiving lots of attention because it shows an improvement on the overall results, this approach has become the latest method that Wieschollek and Lersch used to test on FMD dataset and on a new dataset called Google Material Dataset with around 10000 images divided into 10 material categories. Their work has been able to outperform earlier results found on FMD when natural images are used.

1.2. Problem Statement

The project aims to implement Deep learning Algorithms such as YOLO and CNN for object detection in field of traffic management. The objects are considered as the different vehicles, number plates and humans. The dataset is so designed with specific objects which will be recognized by the camera and result will be shown within seconds using particular algorithm. This application can help for better traffic analysis and can be made useful for building up a “Traffic Analysis System”. This can help for public Security also.

1.3. Objective

This project builds on countless state-of-the-art methods that have been developed over the last years to compare and analyse the best accurate feature extractor when presented with material classification and provides a state-of-the-art evaluation on the best concepts and effective ideas for the task of classifying materials.

My research aim is in developing a better understanding of the best adapted techniques to the task of material classification, performing a thorough evaluation of a range of nine CNN architectures including latest models as well as comparing and analyse the factors that lead to structuring a good dataset for material classification and how it affects the overall system. The focus for this goal is to find the high-level categories in the input images in at least seven common categories – ceramic, fabric, glass, metal, paper, plastic and wood – across all databases. In addition to this topic, it is examined whether by applying transfer learning between the layers of each CNN architecture, the system’s accuracy will produce any noticeable effects on the gathered results. The work presented in emphasizes similar method of comparing CNN architectures. In comparison with research contribution reported in which focuses on recognizing various object categories in an image and using data augmentations, current research aims at providing a complex evaluation of the latest methods on material datasets in computer vision.

The contribution for this research topic is structured in running a large number of experimental tests on countless training and testing datasets to further expand the understanding to which technique is best for material classification and observe whether transfer learning manages to improve results on the current chosen databases. As the current system will return a ranked sequence of scores and documents, there is a need for a good unit

to compute the precision-recall curve and results. A good classifier can rank material images at the top of the returned list. The main performance unit regarding precision-call is called average precision (AP) and it is going to be used on three of the four selected datasets. Compared to just computing and analyse the precision-recall, the average precision offers a simpler way since it is returning a single number that outlines the performance of the classifier and is computed as:

$$AP = \sum \mathbf{P(k)}\ \Delta r(k)\mathbf{(k=1\ to\ n)}$$

where the sum is the precision at a cut off of k images multiplied by the change in the recall. If any slight change it is made in the ranking, it does not affect the score very much, which makes this unit stable and preferred by various researchers. This unit will not be able to be used on the last dataset since each category has a distinct number of photographs per class and the outcomes will be interpreted under specific conditions.

This report starts by introducing the pipeline used for the system, including the several types of visual models and the basic linear classifier used for the supervised learning task. Following this, a series of approaches are described to define the feature extractor hierarchy that is concluded with the results and their analysis. Finally, future directions related to this work are discussed.

Proposed Work

In this paper, we define an effective solution to the computer vision problems using deep learning and GPU techniques. Computer vision provides automatic extraction and analysis of the worthwhile images from a single or a sequence of images. An efficient solution to the computer vision problems are provided in this section.

Overview of Proposed Method

The proposed method works as a combination of deep learning and machine learning approaches. The concept of deep learning is used for object extraction process and to detect and classify the objects machine learning classification algorithms are used.The objective of the work is to automatically detect and classify the images from videos. Here the technique of Convolutional Neural Network (CNN) is used for feature extraction processes. The CNN consists of a series of a convolutional layer with Rectified linear units (ReLu) and max-pooling layers. In addition to this layers, there exist three fully connected layers attached to it. Here CNN is not an original classifier, but it is repurposed to solve different classification problems.

First, the process of image pre-processing is made to remove unwanted noises and features from the input images. Followed by pre-processing of images feature extraction is performed to extract features from the input images. The layers of the CNN captures the input image, then process and extract the features with various layers. The features extracted from each layer of the CNN is combined to form the high-level features. This is achieved through the use of deep learning, and it assists in efficient object recognition process. The features extracted from the previous steps are incorporated in to the CNN to form a pre-trained CNN model. Through the use of the extracted features the Support Vector Machine (SVM) is used to classify the instances in to their corresponding classes. The proposed method make use of cross linear SVM training model for training purpose and it increases the computation speed. Once the training process is completed the SVM classifier is passed with the test dataset for classification processes. Upon the successful completion of the testing processes, the entire workflow is combined to solve the computer vision problem. And the test is performed on a real-time video. The proposed method first classifies the training datasets and then performs the classification process across the test data sets (real time videos).

In this manner, the proposed method achieves its intended objectives and solves the computer vision problems through object detection and extraction using deep learning techniques. The workflow of the proposed model is clearly illustrated in figure 1.

Proposed CNN model

2.4.1.Train the Data

During this stage a wide range of image dataset is imported in to the system. The images collected from various sources are given as an input for the training purposes. Through the use of these images a training model is created and incorporated into CNN model.

2.4.2. Image Pre-Processing

During this process colour images are converted in to grey scale images. Noises from the images are removed and image pixel intensities are extended. A 3x3 mask operates the pixel neighbourhoods. Every time the image is invoked from the image data store a function is invoked to perform the pre-processing process.

2.4.3. Feature Extraction:

Every layer of the CNN produces the response or an activation to the input image. However, only a few layers of the CNN are suitable for the feature extraction process. Feature extraction extracts feature such as edges and blobs for the classification purpose. The beginning layers of the CNN captures the basic image features such as edges and blobs. The network filter weights at the first convolution layer performs

this process. During this stage, it learns the filters for capturing blob and edge features. Next, the network weights from the second convolutional layers are scaled and resized for the visualization process. Thus, the features learned from the previous layers are captured by the deeper network layers, and the deeper layers combine the extracted features of the previous layers and produce high-level image features. Thus, the higher-level features act as the better solution for object recognition processes as it combines the primitive features to form the richer image representation. The process of feature extraction can be easily performed from deep layers through activation functions. Since there exist numerous deep network layers the selection of the layer before the classification layer forms the better solution. The activation function is designed with the GPU support, and higher GPU configuration provides better results. Thus, feature extraction using CNN forms the most suitable solution for the image recognition processes.

Through the use of the extracted features, the objects can be easily detected and classified. The process of object detection involves the extraction of images from the real-world scenarios and providing it as an input to the classification layer. The proposed algorithm make use of the optical flow for the object detection purposes. It utilizes the pixels of the videos from one frame to another for object detection processes. Next, the moving pixels are separated from the image region to analyse and remove the noisy pixels from the video frames. In this manner, the proposed algorithm performs the feature extraction processes. The feature extraction process dimensionally reduces the unwanted features from the input images that optimizes the data fed to the classifier and reduces computational overheads.

A, Grey Scale Conversion

The image of the input is converted into grey scale format: - To convert an RGB image to grayscale, system have to use the RGB2GRAY command from the Image Processing Toolbox. We will use the luminous grayscale formula from the various available ones because it is the most efficient.

- Luminous **Greyscale**: - (0.21Red+0.71Blue+0.02Green)

B. Characteristics Extraction

In image processing, feature extraction starts from an initial set of measured data and builds derived values. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features. Feature extraction involves reducing the amount of resources required to describe a large set of data. In this system we have two algorithms for feature extraction. First is the bags of features and second is the faster R-CNN itself. Feature extraction is very important in the sense of selecting the correct features for a specific kind or set of images.

C. Training

Extensive training of the system is carried out on over more than 20 different objects. The training dataset was taken from the ImageNet dataset which is a very diverse and highly rated open source dataset.

The system was trained with over more than 25 convolution layers designed specifically for optimum performance.

D. Classification

The key features of the image and that which the system is trained on are compared. After comparing the key feature, the system shows the result by identifying the object in the image.

System Architecture

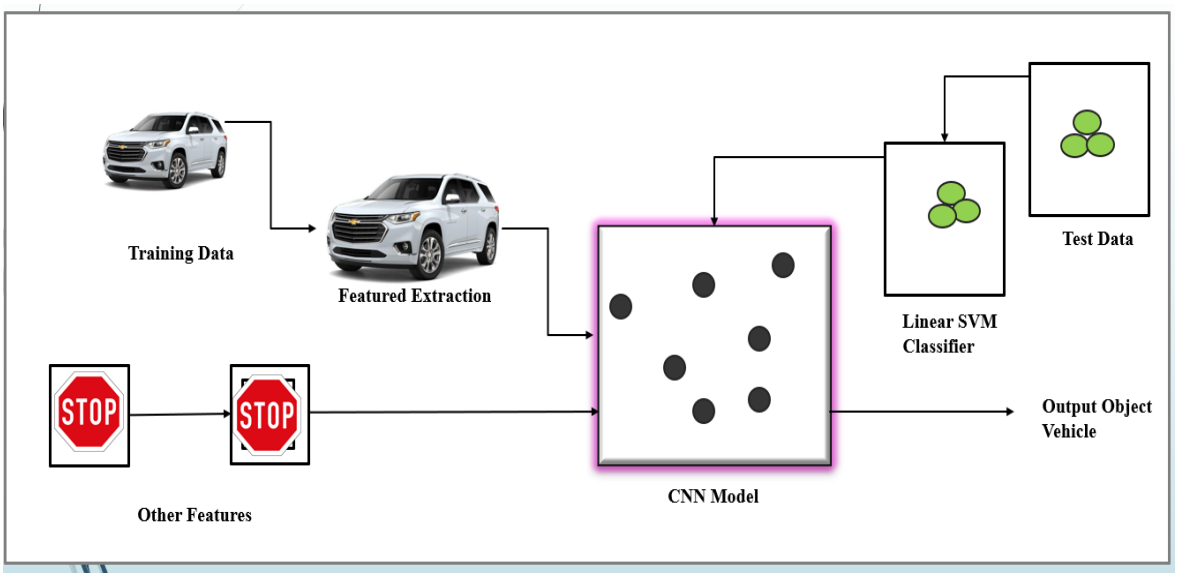

Fig 4.1 System Architecture

We pass an image to the network, and it is then sent through various convolutions and pooling layers. Finally, we get the output in the form of the object's class. Below is a succinct summary of the steps followed in RCNN to detect objects:

1. We first take a pre-trained convolutional neural network.
2. Then, this model is retrained. We train the last layer of the network based on the number of classes that need to be detected.
3. The third step is to get the Region of Interest for each image. We then reshape all these regions so that they can match the CNN input size.
4. After getting the regions, we train SVM to classify objects and background. For each class, we train one binary SVM.
5. Finally, we train a linear regression model to generate tighter bounding boxes for each identified object in the image.

Outcomes and Discussion

When a system dataset is supplied to a system, it will apply different image processing techniques to it. Accuracy has improved a lot. In the sample set more than 90% and more than 75% have been obtained with a sample set of 10 classes for 20 classes. Therefore, we can conclude that increasing the sample set size reduces the accuracy. We can cope with this by training the system on a large dataset.

Parallel computing toolbox, statistics and machine learning toolbox, computer vision toolbox, image acquisition toolbox and neural networks toolbox. Further, the implementation process requires the Matlab compatible CUDA-enable GPU. The experimental analysis of the proposed algorithm is made with the pet recognition application. For pet detection we made use of two labels 'Dog' and 'Cat.' An object detected can be categorized into either of this classes. A massive collection of the pet images is extracted from the internet and other resources to create the training data store. First, the CNN model is trained with numerous common objects using the imagenet dataset. The imagenet dataset contains features of more than 1000 objects. Once the CNN is trained with the features of the imagenet dataset, it performs the process of feature extraction in an appropriate manner. Once the CNN model is built with the trained model, the multi-class linear SVM is invoked to perform the classification process.

First, the process of feature extraction is made using the CNN. The convolutional layer performs the classification processes using activation method. The activations are arranged in a columnar manner. This assists the linear SVM in the classification process. Next, the classification process is made with the training data store. It consists around 1500 images. Since the data store contains 1500 images, the batch size has been set to 750images. This can be reduced with lesser GPU capabilities. As a result of the classification with the training dataset first, the proposed algorithm generates a 2x2 table with the labels and the count. Where the label denotes the object class and count denotes the number of objects belonging to the class. Figure 2 represents the 2x2

Classifier Result of the proposed algorithm Since we make use of the GPU support the time to classify the instances are compared with CPU and the GPU environment, and it is described in figure. It is found from the observation the total time to classify 1328 images in a CPU environment takes around 27 seconds and nearly around 2 seconds in a GPU environment. This is because of the reason that the deep learning technique becomes inefficient when applied

to the larger datasets. Thus the use of GPU provides better results even with larger datasets. A comparison to the CPU and GPU computation measures is clearly illustrated in figure4. From this observations, we conclude that the deep learning solution with

GPU support forms the most efficient solution for real-time video analytics purposes.

The proposed algorithm acts as a better solution when it is applied under certain constraints. Table provides a summary on the applicability of the proposed method. Table 1. Summary on Applicability of the proposed algorithm

Training Data 1000 to millions of label images Computation Compute intensive (Requires GPU) Training Time Days to weeks for real problems Model Accuracy High (can overfit to smaller datsets)

Fig 5.1 Result -1

Fig 5.1 Result - 2

5.2 Conclusion and Future Work

Due to the variability of the materials, recognizing materials using supervised learning is a very challenging task that has received lots of attention in the last decades. This project targets specifically object classification and presents an accurate and observational evaluation of nine distinct CNN models on four varied datasets. Since segmentation and image understanding are some of the fundamental challenges computer vision systems attempt to tackle, this project took a further look at approaches like patch segmentation and transfer learning and how they affect the way the features are learned by the network at different layers.

Through the experimental tests, the real-world scene understanding has been examined by looking at the contextual modelling between the diversified components of the created system. The pipeline consists of the training and testing sets that are fed as inputs to the pre-trained network on the ImageNet dataset. The network will then extract the features in a feature

vector that is fed to the classifier which will compute the score map. The mean average precision will rank each image in the dataset and output the results.

After the system is provided with the image it will apply various image processing techniques on it. The performance in the terms of accuracy has greatly improved. Accuracy of more than 90% was achieved with the sample set of 10 classes and more than 75% for more than 20 classes in the sample set. So we can this normal trend which is to be expected i.e. as the size of sample increases the accuracy decreases. We can counter that to a certain extent by training the system on an extremely large dataset.

Automatic machines including car has given a boom exposure to IOT field of technology. Adding automatic traffic analysis system to the car can be a great milestone to be achieved. Hence, analyzing (traffic signals, animals, persons) in front of car which can provide sensor break instructions to the car is the future target of the project.