International Journal of Interdisciplinary and Multidisciplinary Research (IJIMR)

ISSN 2456–4567

Autonomous Driving Based on Deep Learning Image Recognition

Muhammad Afzal Nazim, **Bakhtawar Sarfraz**
Department of Computer Science
Government College University Faisalabad, Layyah Campus

Abstract

In the years before 2010, researchers used the local picture features that they had discovered with hybrid machine learning techniques to solve issues with image identification. Some deep learning methods for picture identification have, however, been created and tested since the year 2010. The methods that utilize deep learning to identify pictures outperform strategies that were used before deep learning was introduced in generic object recognition competitions by a considerable margin. Because of these developments, this article describes how deep learning is being used in the field of picture recognition, in addition to the most current achievements in autonomous deep learning driving.

1.Introduction

With the advent of general-purpose computers in the late 1990s, large quantities of data could be processed at breakneck rates for the first time. Traditional methods included extracting the vector from a picture and then using a machine learning technique to identify the visual object in question. Controlled machine learning, on the other hand, does not need the establishment of new rules, in contrast to regulatory methods. An enormous number of class-labeled training examples are used to instruct the computer instead. As a consequence, it is feasible to do flexible picture recognition. During the years 2000 to 2005, a significant amount of research was conducted on the application of artisanal features, such as the scalable invariant feature transformation and the histogram of oriented gradients, as local picture characteristics that were developed on the basis of the researchers' expertise. In the case of facial identification, the combination of local imagery with machine learning has resulted in an improvement in the performance of real-world image recognition applications. When deep learning was first introduced in late 2010, it was considered as a machine learning method for the extraction of functions, rather than as a standalone technique. It is not always the greatest option to use a manufactured feature, since it is designed to extract and express distinctive characteristics using a technique that was specially created based on the researchers' experience in the process of its development. Deep learning is an image recognition method that, by learning from examples, has the potential to automate the extraction process. In spite of the fact that picture identification, which is required for autonomous driving, is still in the early stages of study, deep learning has produced outstanding results in generic object recognition competitions. This article covers the growing trend toward deep learning-based autonomous driving, as well as the challenges that come with it, as well as the application of deep learning to every task in image identification, as well as the problems that come with it.

2.Problem setting in image recognition

Performing generic object recognition issues in conventional machine learning from an input image is challenging because of the nature of the task. In order to solve this problem, the tasks for image identification, picture classification, object detection, scene understanding, and specific object reconnaissance should be separated, as shown in Fig.1. The definitions and procedures for each task are described in more detail in the following sections.

2.1 Image verification

The aim of the picture verification is to assess if the item in the photograph conforms to the pattern of reference. During the image verification process, the distance between the reference pattern vectors and the input image feature vectors is computed through the reference pattern feature vectors.

If the distance value between the two pictures is less than a specified value, the images are the same. Otherwise, the images are the same. One job that requires determining whether or not one person is the same as another comprises fingerprinting, face recognition, and identification. Deep learning is used to address the issue of identification by creating a loss function (the threefold loss feature) which evaluates the distance between two photographs of the same person and the distance between two pictures of the same person.

2.2 Object detection

item detection is the problem of identifying an image that corresponds to a certain category and finding the item in question. This project will offer hands-on experience in the areas of face recognition and pedestrian identification, among other things. Human-like characteristics and AdaBoost are utilized to identify the face, while HOG features, vector support machines, and vector support machines are used for pedestrian identification. Traditional machine learning, for example, identifies objects by training 2-class classifiers that correspond to different categories and then scanning the picture using a raster image recognition algorithm. When using deep learning-based object detection, it is possible to do multiclass object identification using a single network that targets several categories at the same time.

2.3 Image classification

The task is to determine which of the preset categories of an item in an image is referred to as a categorization of the image. Traditional Machine Learning uses a technique called bag of characteristics, which is short for a bag of characteristics.

A histogram quantifies the local qualities of the image and shows the features of the whole image. Very contrast, deep learning fits in well with picture classification problems, and by defeating people in 2015 it earned notoriety in a 1000-class image classification test carried out in a controlled setting.

2.4 Scene understanding

Scenario understanding refers to the difficulty of comprehending the way scenes are organized in a movie. In addition, traditional machine learning has had difficulties coping with semantic segmentation, a procedure in which item categories have been identified in each pixel of a digital image. As a result, it has long been considered one of the most challenging computer vision issues; nevertheless, new research shows that the problem may be resolved by means of a deep learning process.

2.5 Specific object recognition

The phrase used for describing the issue of locating a particular item is specific object recognition. Specific object recognition is described as a general object identifying issue subtask and is achieved by associating characteristics with properly named items (see Figure 1). The recognition of feature points with the help of a computer makes it possible to identify certain objects.

SIFT and the voting technique based on the computation of the distance between reference pattern features are all explained here in-depth. Although machine training has not been specifically used in this study, LIFT, which was created in 2016 and enhances performance via learning and replacement of each SIFT step with deep learning, has been employed to boost performance.

3.Deep learning-based image recognition

Before profound learning, image identification was not exact since image characteristics were removed and represented using a handmade feature, an algorithm that researchers developed based on their expertise and experience. Fig 2 shows a neural convolution network, a thorough learning method for categorization and extraction of training data presented in the figure. This chapter offers a summary of CNN, concentrates on part detection and scene, and discusses its application to the identification of images, as well as current research and progress.

3.1 Convolutional Neural Network (CNN)

A CNN is a kind of neural network trained to learn new information.When the kernel in the picture is converted, the function map corresponding to the kernel is computed using CNN as illustrated in Fig.3. Since there are numerous kernels, feature maps that correspond with different types of kernels may be computed. The characteristic map that is pooled reduces the overall size of the characteristic map by one. Therefore, a processing technique is capable of accommodating little changes in the input picture, such as minor translation and rotation. The data characteristic map is created via a series of convolution and pooling operations that are repeated over and again. It is input into fully linked layers, which then output the probability of each class depending on the feature map that was previously collected. In this example, both the input layer and the output layer have a network structure, with image units and a number of classes in the input layer and a number of classes in the output layer, respectively.

The backpropagation method is utilized to change the network parameters, and CNN training is carried out throughout this process. The parameters of a CNN are the kernel of the coevolutionary layer and the weighting parameters of all connected layers, which are defined as follows: A diagram of the backpropagation method's process flow is shown in Figure 3. Initial training data are fed into the network, and existing parameters are used to make predictions based on the training data. After the prediction and the training label have been calculated, the quantity of data that should be sent from each parameter's output to its input layers is calculated, and the network is updated as a result of this computationIt is necessary to perform these processes many times to get the right parameters that will enable the computer to correctly identify images.

3.2 Benefits of CNN to conventional machine learning

As shown in Fig. 4, the figure shows different kernel visualization examples from the first overlay layer of Alex Net developed for the ILSVRC classification 1000 items. It consists of 5 convolution layers and three fully linked layers, with the output layer of 10,000 units which corresponds to the number of classes in the input layer. The Alex Net automatically acquires a number of filters that extract information on the directional edge, texture, and color, as shown in the following picture. To evaluate how efficient the CNN filter is local image feature, in the identification test HOG was compared with the HOG. HOG has an error rate of 8%, while CNN filters have an error detection rate of 3%. Although CNN kernels utilized by Alex Net were not particularly trained for the identification of humans, the precision of detection increased beyond the conventional handmade HOG feature previously employed.

Fig. 1 shows how the CNN can not only categorize pictures but also identify objects and semantically describe them by creating an output layer that is suitable for each image recognition task, as shown inFig. 5. Using the probability of output layer and the detection area of each grid, for example, a network structure capable of detecting objects may be constructed.The output layer should be designed using semantic segmentation such that it provides the class probability for each pixel. For this reason, convergence and pooling layers may be utilized as shared modules like other processing layers can be used. However, it was required to create local image characteristics for each job using traditional machine learning techniques and to integrate this with machine learning before it could be utilized in machine learning. Due to the flexibility of CNN, it can be used for many different tasks by changing the topology of the network, which is especially useful in the area of image identification.

3.3Uses of CNN to object detection task

In traditional machine-based item identification, a raster scan is used to identify the object using two classifiers. In this instance, the fact that the aspect ratio of the item is constant implies that just one category of object identification is learned as a consequence of the positive sample. In contrast, object detection using CNN detects object proposal areas with a range of features, and multiclass object detection may be carried out using the Region Proposal method that uses CNN to do a multiclass classification for each region identified. Faster R CNN presents the Regional Network for Proposals, which concurrently identifies object candidates and object classes in these regions, as shown in Fig. 1. For a feature map to be generated, convolution processing must first be carried out on the whole picture being analyzed. When detecting an object in the RPN, a raster scan is used to find the item from the detection window on the resulting feature map. The raster scanner anchors are targeted areas in which detection windows are applied to target regions known as anchors ask-numbers of forms. RPN provides a score showing how similar two items are as well as the position on the input picture in response to the area provided for the anchor. Moreover, when RPN determines that the area indicated by the anchor is an object, they transmit the information for object identification to an all-connected network to further validate the findings. A rectangle split by class number generates the output layer whose unit is the sum of class numbers and (x,y,h) classes divided by one rectangle. By utilizing these region proposal techniques, several classes of objects with different aspect ratios may now be identified.

This year, the single-shot technique, a novel method for identifying multiclass objects, was shown. The entire picture is included in the CNN algorithm to detect a huge number of objects without raster scanning the image. One example of this is shown in Figure 1. In step 6, the YOLO method creates an item rectangle and category in each local area split by a seven-by-seven grid into seven equal parts. Initially, feature maps are produced by converting and pooling input pictures, followed by additional processing. The location I j of each canal is used to create a feature map of the image input and this feature map is fed into fully linked layers. In addition to the position, size, and confidence of the two object rectangles, each score for each item category at each grid point includes all of the output data collected via fully connected layers as well as the location, size, and confidence of the two object rectangles. Therefore, the unit measure of the output layer is the number (1470), which is obtained by multiplying the number of categories (20 categories) by the number of grids (7 7) and then adding two rectangles ((x, y) and reliability) in the locations of the categories and grids, as well as in the size and reliability. As long as YOLO is utilized for object identification purposes, the detection of object area candidates such as Faster R-CNN will not be possible in real-time. Figure 7 depicts an example of a YOLO-based multi-class identification system.

3.4 Uses of CNN to semantic segmentation

For a very long period, computer scientists have been confronting the challenging problem of synanthropic segmentation, which has been extremely unpleasant. Conventional methods to machine learning beat deep learning strategies suggested for other tasks based on the research findings. Deep learning approaches have been proven to be considerably lower than conventional methods. Using the assistance of a completely co-evolutionary network, end-to-end learning can be achieved and segmentation results can only be provided with a CNN as a starting point. Figure 8 shows the Federal Communications Commission's organizational structure. According to the FCN, the architecture of the specific network does not contain a layer that is fully linked with the rest of the network. Using a coevolutionary layer and the pooling layer on the same input picture several times, the size and complexity of the resulting feature map may be decreased substantially by applying both layers to the same image many times. To guarantee that the final map layer matches the original photograph, the last layer of the feature map is enlarged 32 times after the transformation is complete and is then scaled down to the same size as the original picture. This particular method is referred to in the scientific world as "deconvolution," which is its technical term. It is the last layer of the algorithm that produces probability maps for each class, made in the final algorithm layer, which is the final algorithm layer. The probability map must first be trained before the end-to-end segmentation model can be used to compute the likelihood of every class in each pixel. The output unit of the end-to-end segmentation model is used to compute the probability of each class in each pixel using the output unit of the segmentation model after training the probability map (w h number of classes). As it is closer to the input layer of the network than the first layer feature map, the middle layer feature map of CNN offers far more detailed information than the first layer feature map of CNN. The middle layer feature map of CNN offers more detailed information than the first layer of the CNN feature map. However, because both pieces of information are combined during the bundling process, it is not feasible to distinguish between the two types of data. With the aim of broadening the coverage of this feature map, you will see some early outcomes of segmentation from your work. This merger combines the intermediate layer feature map with the base layer feature map to obtain high precision via the combined data. Functional maps may be integrated into the core of the network via the use of FCN performance processing, which is a network feature. The convolution segmenting process combines mid-function mappings in the direction of the channel to produce segmentation outputs identical to the source picture. This is accomplished via the use of mid-feature mappings towards the channel.

PSP Net may gather data at a variety of scales while simultaneously extending the encoder function map. It is achieved via the use of the Pyramid pooling unit, which expands as the input is gathered in various forms and sizes. The feature maps of the encoder are grouped with numbers 1, 2, 3, 3, 6, 6, and 6 in the Pyramid-style pooling module, while the horizontal and vertical widths of the original designs are decreased to 1/8 on the encoder side. The feature maps are grouped with the encoder numbers 1, 2, 3, 3, 6, 6, and 6 using a pyramid-style pooling module. Every feature map is then subjected to a reconfiguration procedure to further improve its overall quality. The process of convolution is finished when all function maps have been expanded and linked to the same size as probability maps have been produced for each class. The PSP Net was the best scene parsing technique prize winner at the 2016 International Scene Parsing Competition in New York and was the top prize winner in the scene parsing category. With the Cityscapes Dataset, recorded by a dashboard camera, a high degree of accuracy may be obtained in mapping. The findings of research on the net semantic segmentation of the PSP networks conducted using the PSP networks are displayed in this figure.

3.5 CNN for ADAS application

System intelligence may be included in an ADAS via the use of machine learning methods (Advanced Driving Assistance System)

In order to provide the driver, the latest information from sonar, radar, and cameras, advanced driver assistance systems (ADAS) have been created. However, although radar and sonar are often employed for long-range detection, in the case of pedestrian identification, lane detection, and redundant item detection, CNN system-based may possibly play a major role.

The three main components of independent driving, perception, planning, and control may be divided into three groups. Based on our perception, we may comprehend our surroundings, such as the identification of obstacles, the identification of road markings, and the categorization of items based on their semantic labels. Localization refers to an autonomous vehicle's capacity to determine its location with respect to its environment. Planning involves decision-making to accomplish the vehicular objectives, which usually include transporting the vehicle from a place of departure to a destination while avoiding barriers and maximizing the route's efficiency Finally, control refers to the vehicle's ability to carry out the duties that have been assigned to it during the whole planning phase of the project. Because it can handle objects of a wide variety of shapes and sizes, CNN-based object recognition is very effective for perception. As a result, by referring to pixels that have been classified as roadways, semantic segmentation provides important information for decision-making in planning in order to avoid bottlenecks.

4. Autonomous deep learning driving

This chapter discusses profound learning models and future problems such as end-to-end self-driving learning, which may directly derive the value of control for cars from the input picture as an example of an application of deep autonomous driving education.

4.1 Endtoend based autonomous driving

According to most autonomous driving studies, a dashboard as well as light detection and the range technology are used to understand the environment surrounding the car, an adequate position in movement planning is established, and the control value of the vehicle is calculated. These three methods became increasingly common in self-employed driving and are currently utilized to assist cars to comprehend their surroundings via the deep-seated object identification and semantic segmentation techniques described in Chapter 3.In contrast to the previous advances in CNN research as the field develops a learning technique was suggested for inferring the value of the vehicle directly from the input picture. In this way, the network must learn from pictures taken by a dashboard camera being driven by a human person. The network also teaches how to learn from the control settings for each frame captured on the dashboard camera used to train the network. In particular, a self-supporting control system based on end-to-end learning simplifies system set-up since CNN automatically and reliably learns without specific environmental or motion planning information.

This is the objective of Bojarski and others' autonomous driving methods, which includes pictures from the dashboard in a CNN and a directory of steering angles. Numerous studies have been conducted including the creation of a methodology for analyzing the temporal structure of a dash-cam video and a method of CNN training using a driven simulation system and use the trained network to drive a car in a real environment. While the steering angle is regulated using these techniques, the throttle is controlled by the driver. This article provides an autonomous driving model, in which both the steering and the throttle and other factors are included in the control value of the vehicle. The bundling technique utilized for the construction of the network architecture consists of five layers of convergence and three levels of completely linked layers (seeFigure 1). Moreover, because the change of speed in your own vehicle is essential to manage the throttle, the measurement is performed in accordance with your own condition by giving vehicle speed in addition to Dashboard images, the measurement is carried out using the fully linked layer while taking into account one's own state High precision steering and throttle control may be utilized to manage a range of driving situations if this method is applied properly.

4.2 Visual explanation of end-to-end learning

This creates a challenge since CNN-based end-to-end learning is difficult to implement when the basis of the output control is not known. In order to resolve this problem, researchers are investigating a technique of selecting decisional criteria that is comprehensible to humans (such as moving the steering wheel to the left or right and stepping on the brakes).

The use of a visual explanation is a common way of communication for explaining the objectives of network decision-making and other network-related topics. With the visual explanatory method, a heat map of the region in which the network is concentrated is generated, which represents an attention map of the area where the network is concentrated. It is possible to analyze and understand the reasons for making choices based on the map of attention that we have created for ourselves. For effective visual explanatory reasons, a number of methods have been proposed in the field of computer vision to create an attention map that is clearer and more informative. Feature maps from the last coevolutionary layer of the network are weighted and averaged to generate care maps, which are then used to construct care maps. Gradient-weighted class activation mapping is a common method that creates an attention map by using gradient values calculated during the rear propagation phase of the learning process.This technique is popular for an exhaustive study of CNN's since it can be applied on any network.Figure 12 illustrates the CAM and Grad-CAM attention maps.

Several visual explanation approaches have been developed for general imaging tasksand visual clarification methods in autonomous driving have also been proposed. Several visual explanation methods have been developed for broad picture identification tasks. The visual backpackintegrates the characteristics charts for each coevolutionary layer in a single map, which is subsequently shown to constitute the intermediate value in coevolutionary neural networks. Using this technique, we can observe which sections of the network react more to the image that is supplied into the network. Fig. 1 shows a network of a branch of attention comparable to the network reported in reference. CNNs are divided into two sections: a function extractor and a regression branch. The function extractor is the initial component of the CNN. A new branch of attention is added to CNN in order to visually display the map on the screen. Direction and throttle control may be created in a range of situations by providing vehicle speed in totally connected layers and by learning from the beginning to the end of every branch of the Attention Branch network. This method may also be used to generate an attention map that explains where the control value has been shown in the image in question. Following the Regression in an automated self-contained driving scenario, as illustrated in the attention map is given in an automatic self-contained driving scenario based on the Regression of the Attention Branch.

The letters S and T, as shown in the picture, are used to denote the controls for the steering wheel and for the grip. Consider the situation below: A road curves to the right, as shown in Fig 10, and the vehicle shows strong responses to both the central line and its directional performance to suggest that it is moving in the correct direction. This is presented in a very powerful way, on the other hand. While the road curves to the left and the steering output indicate that the road is in the left direction, the map of attention reacts strongly to the white line on the right. In this case, the directional output value is a negative integer that shows the direction to the left of the current position. The Centreline of the road and the location of the lane are then checked to establish the steering value to be utilized once the map is displayed in this way. The car also stops at the location, as indicated in the picture. The study showed that the attention map reacted considerably to the brake light of the car in only 12 seconds, which is a significant amount of time in a very short time (c). There will thus be no throttle output, indicating that neither the accelerator nor the brake is utilized. The status of the car facing you is taken into account when deciding the setting of your throttle, which is not surprising.In addition, the situation illustrated in Fig. 1 that shows night travel as the road shape ahead is uncertain, the attention map reacts significantly to the vehicle. 12 (d) illustrates a scenario in which a vehicle is being followed and the map indicates a strong reaction to the automobile ahead since it is not known in advance of a road shape ahead. Using the output of the attention map, the rationale behind a choice may be graphically explained using the attention map.

4.3 Future challenges

Engineers and academics may easily examine and comprehend, as the consequence of the visual explanations given, the internal state of deep neural networks. Adequate explanations to end-users such as passengers on a self-driving car will be one of the future difficulties. Even when there are no vehicles ahead or on the side of the roads, completely independent driving may change lines suddenly without a warning, leading the driver and any passenger to be worried about why the lanes have changed. Individuals in this situation may use the visualization technique for maps of attention described in Section 4.2 to understand more clearly why they are shifting their lanes. On the other hand, displaying the map in a fully autonomous car is useful only if someone is continually informed of what is going on in the car. In

"Changing to the left lane when a vehicle from the back approaches at speed," a person in an autonomous automobile, i.e., a person who benefits completely from AI, must be advised in writing or voice of decision-making criteria by saying: "Changing from the rear left lane is approaching at speed." In the next step of the development process, the transfer from recognition results and visual explanations to voice explanations will be handled. In oral explanations, it is still challenging to attain adequate precision and flexibility despite many attempts.

Indeed, in the not-too-distant future, such voice explanation services will become redundant. Those who benefit the most from independent driving may find it difficult to accept at first; nevertheless, by repeating their verbal explanations, a feeling of trust in the people impacted will eventually develop. This will result in verbal explanation functions being unnecessary after confidence has been built between autonomous AI driving and individual driving and autonomous driving based on AI will become ubiquitous and widely accepted.

5. Conclusion

In addition to explaining how deep learning is used to address problems in photo identification, this article covers the latest deep learning imaging technologies available at the time of writing (at the time of writing). It is hard and time-consuming to detect the appropriate mapping function from many training data and instructor labels in a time-consuming and difficult field of image recognition technology. Multi-task learning also enables you to address a range of problems simultaneously, which is extremely useful. It's not only the development of deep learning technology for "judgment" and "management" of self-driven vehicles which generate a great deal of enthusiasm, but it's the development of picturesque technologies of "recognition." It also is desirable to move from visual to verbal explanations by integrating natural language processing with in-depth learning and profound learning enhancement in practical applications, as it represents a significant problem in practice to state the reason for deep learning and the output of deeper learning.

References

1. *Distinctive image features from scale-invariant keypoints, D.G. Lowe, Int. J. Compute. Vis. 60 (2004) 91–110.*
2. *Visual explanation by attention branch network , K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Proc. of IEEE Intelligent Vehicles Symposium (9-12 June 2019)*
3. *Learning deep features for discriminative localization, by B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba,*
4. *Realtovirtual domain unification for end-to- end autonomous driving, Y. Luona, L. Xiaodan, W. Tairui, X. Eric, Proc. of European Conference on Computer Vision (2018) 553–570.*
5. *Attention neural baby talk by Y. Mori, H. Fukui, T. Hirakawa, N. Jo, T. Yamashita, H. Fujiyoshi.*
6. *The cityscapes dataset for semantic urban scene understanding, M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, Proc. of IEEE Conference on Computer Vision and Pattern Recognition (27-30 June 2016)*
7. *Muhammad Alam, Jian-Feng Wang, Cong Guangpei, LV Yunrong , Yuanfang Chen , Convolutional Neural Network for the Semantic Segmentation of Remote Sensing Images,link.springer.com*
8. *Rapid object detection using a boosted cascade of simple features, P. Viola, M. Jones, Proc. of IEEE Conference on Computer Vision and Pattern Recognition,.*

Figures

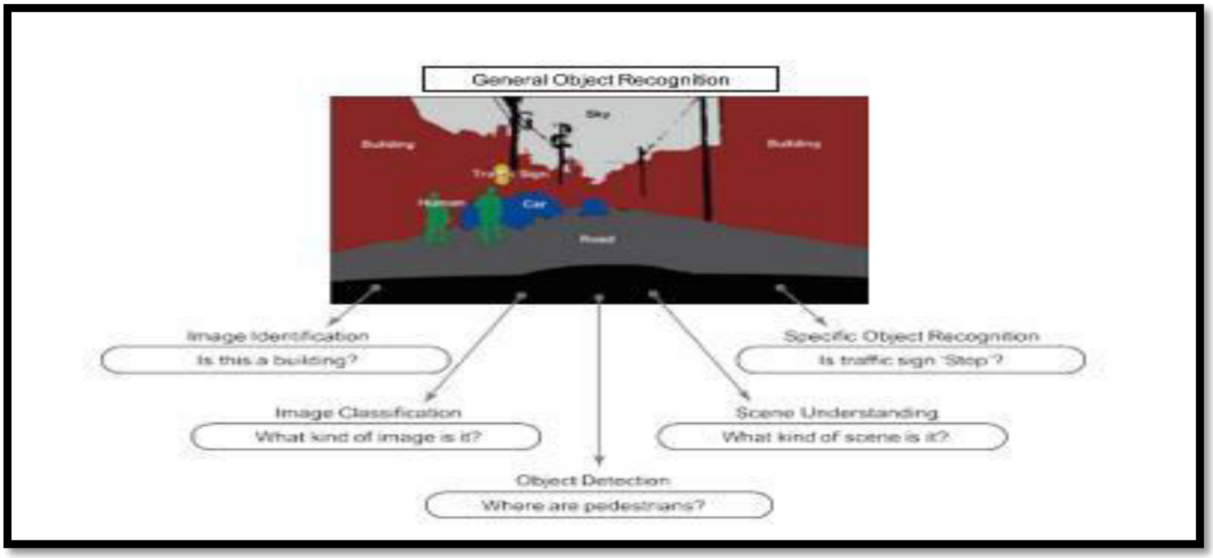

Fig. 1. Part of general object recognition.

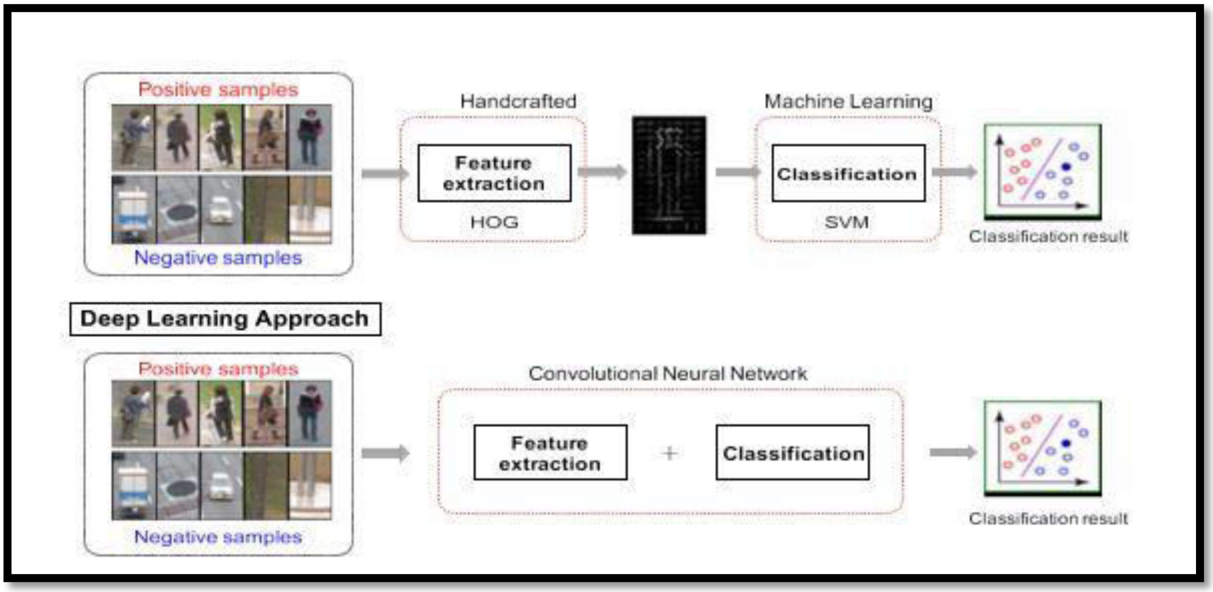

Fig. 2. Conventional deep learning and machine learning.

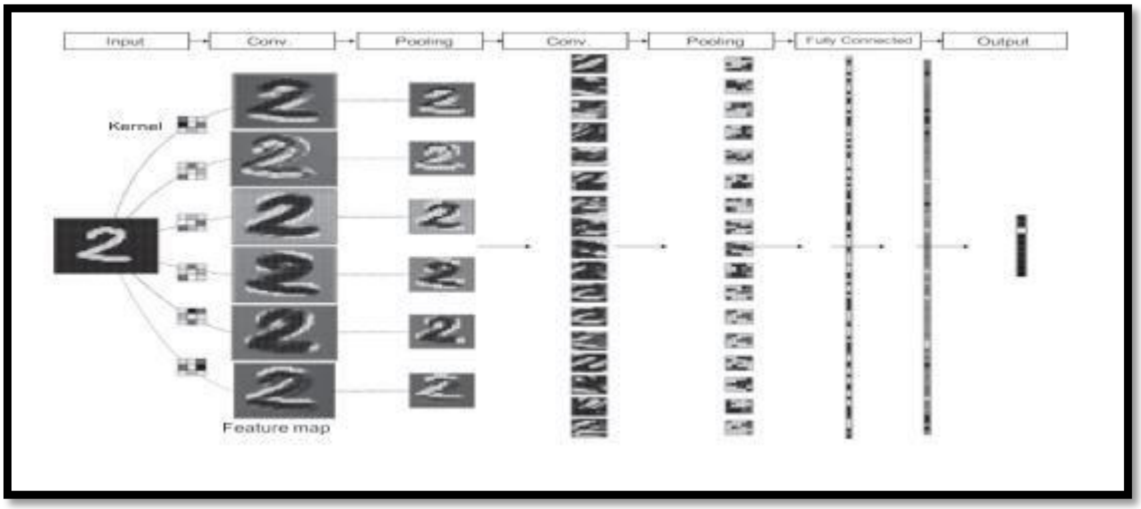

Fig. 3. structure of CNN.

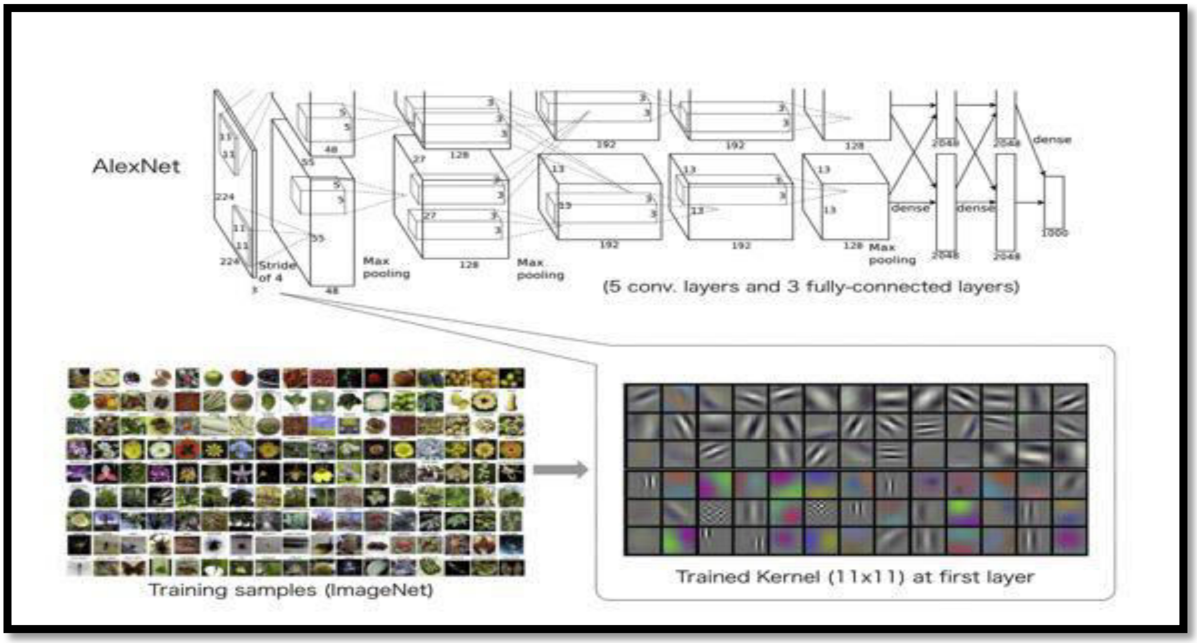
AlexNet
224
224
11
11
Stride of 4
55
48
5
27
128
Max pooling
13
192
128
dense
2048
1000
(5 conv. layers and 3 fully-connected layers)
Training samples (ImageNet)
Trained Kernel (11x11) at first layer

Fig. 4. structure of Kernels and AlexNet.

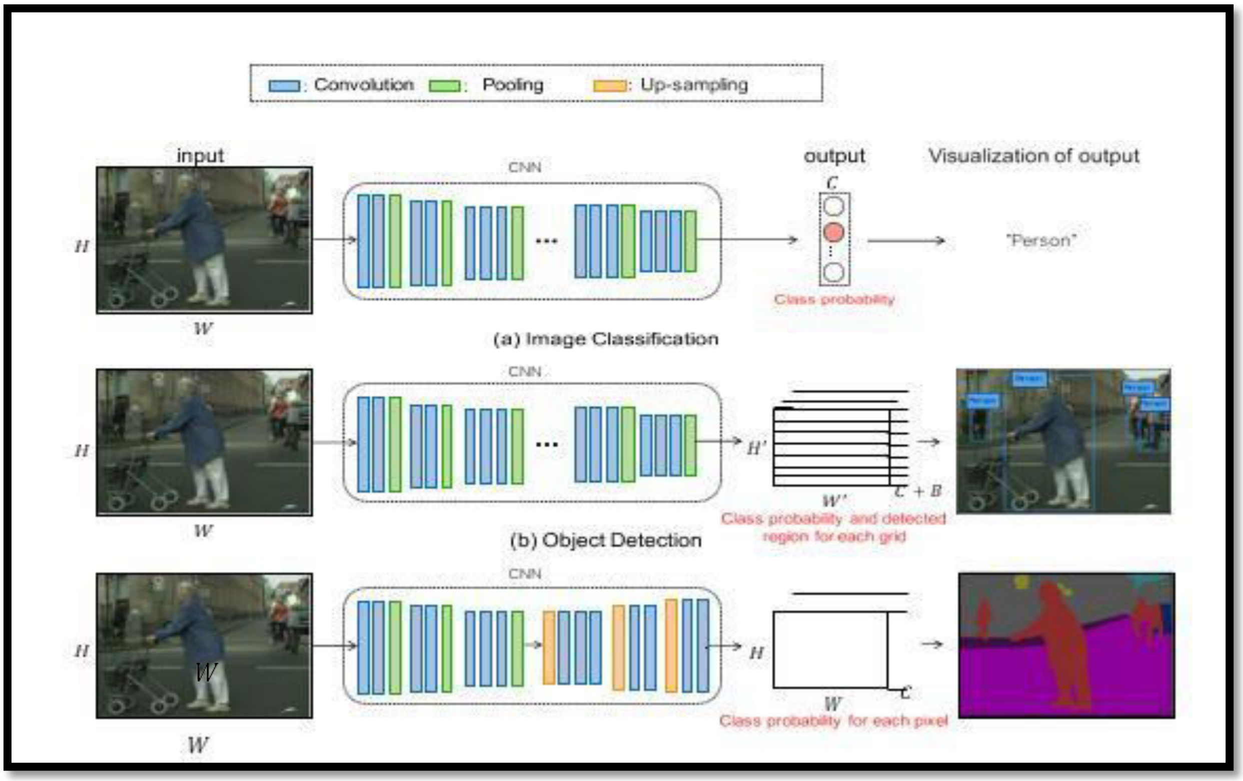

Fig. 5. Application of CNN to image recognition.

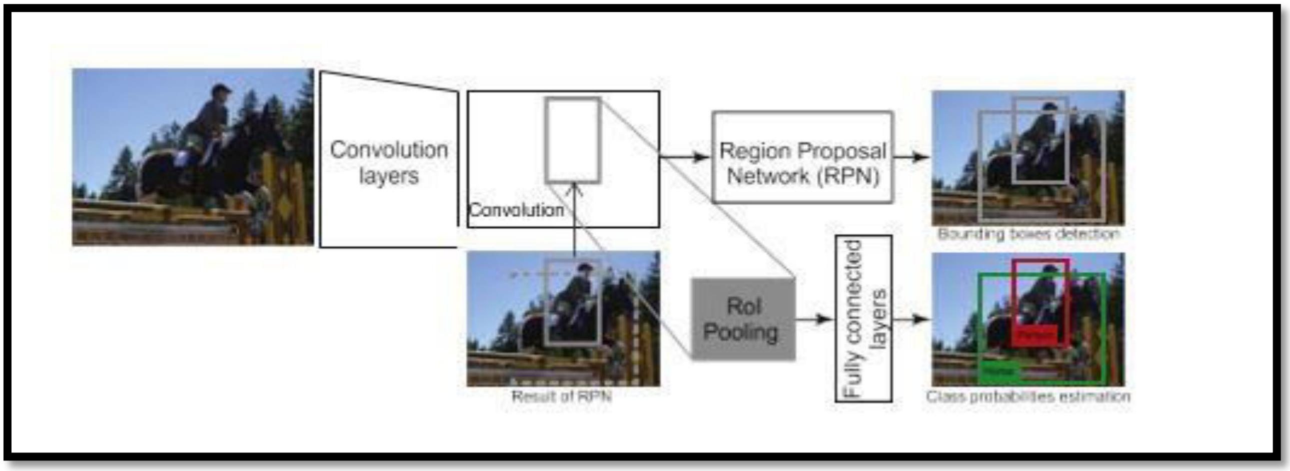

Figure. 6 Faster R CNN structure.

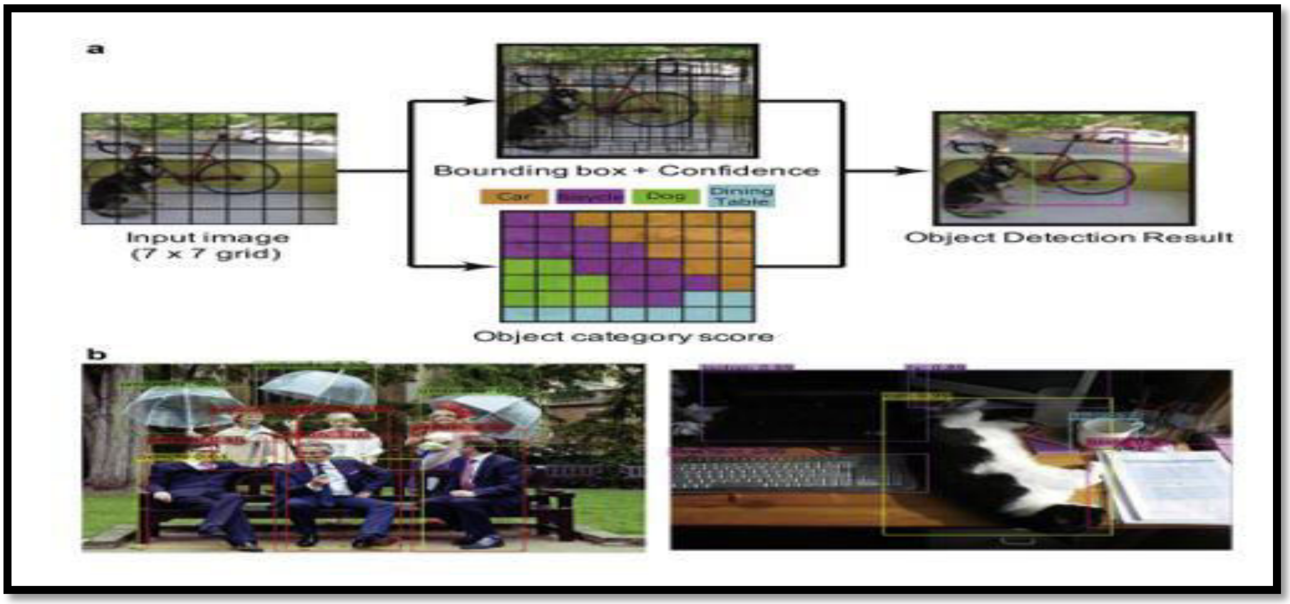

Figure. 7 YOLO structure and multiclass object detection.

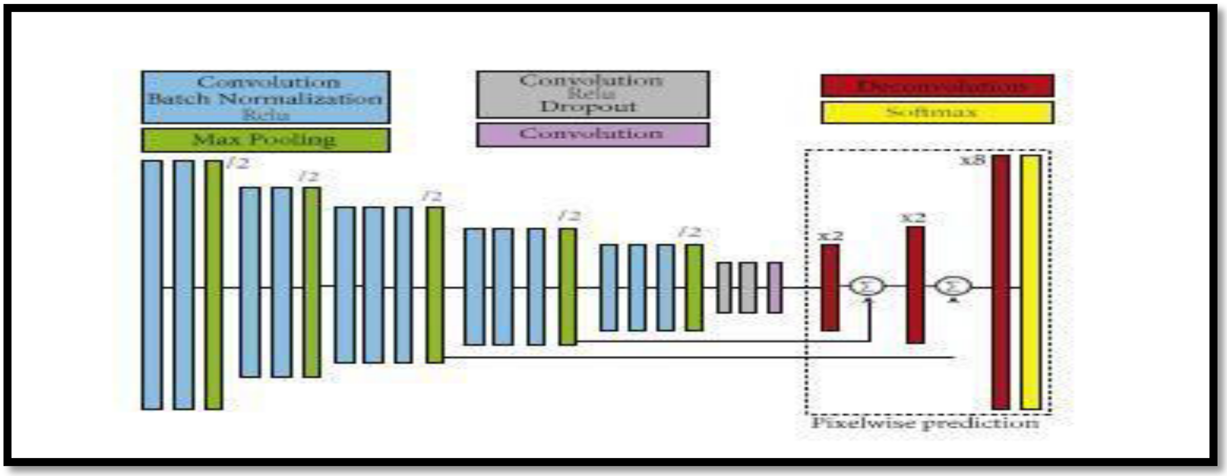

Figure. 8 Fully Convolutional Network Structure.

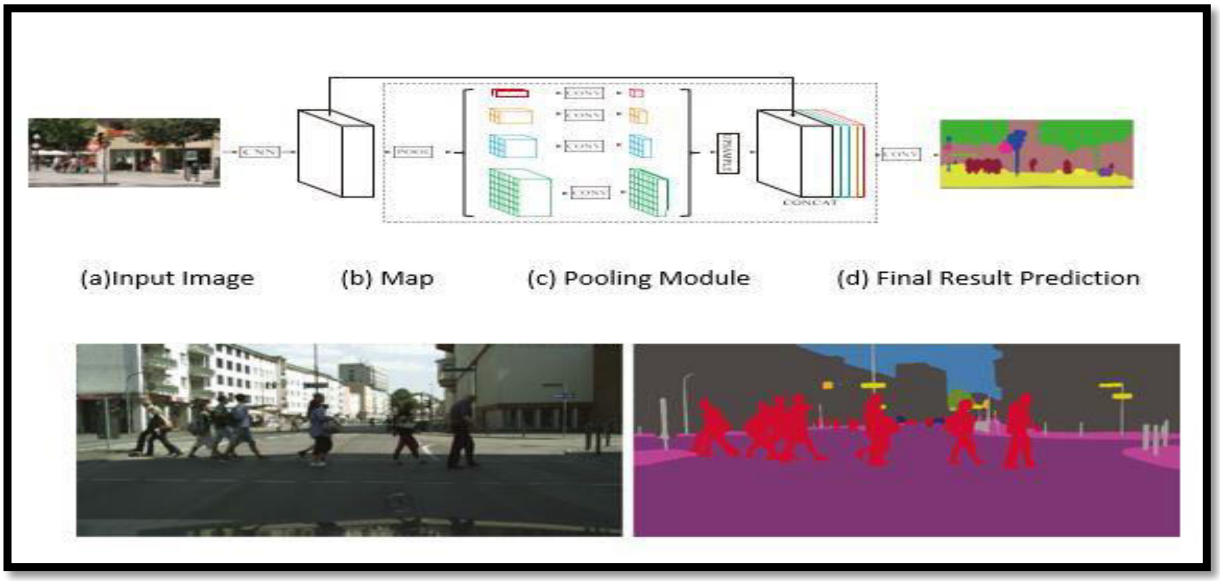

Figure . 9 Example of PSP Net-based Segmentation Results

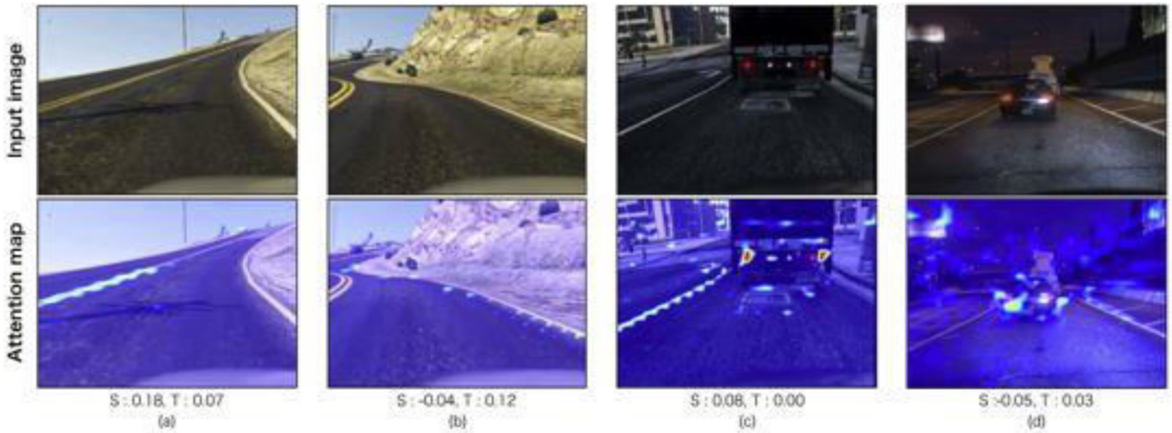

Figure 10. map-basedvisualexplanation