International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022 ISSN: 2348-6600
Reference ID: IJCS-389 PAGE NO: 2657-2667

Paper ID : ICESA029 *Date of Conference: 12.11.2021 &13.11.2021*

International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology

Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

A Deep Learning Perspective of Computer Vision Algorithms

P.Praveena

Department of Computer Application & Information Technology,
Thiagarajar College, Madurai -9. TamilNadu, India.
praveena_caitsf@tcarts.in

Abstract— **The field of computer science that deals with replicating the complex parts of the human visual system and getting the machines to mimic the brain for visual details present in the data is known as 'Computer Vision'. CV is a very popular field that is used in almost all the domains that work with images and videos. It finds itself solving really complex problems in a simple manner and it is easy to train and work with as well. This paper, review the important tasks such as Classification, Segmentation and Object Recognition of CV and their deep learning algorithms which plays vital role on Computer Vision. Deep learning allows computational models of multiple processing layers to learn and represent data with multiple levels of abstraction mimicking how the brain perceives and understands the image.**

***Index Terms*— Computer Vision: YOLO, Segmentation, Object Recognition, Convolutional Neural Network.**

1. Introduction

In this modern era we are flooded with all kinds of photos from Selfies to Landscape images today. A report by Internet Trends says, people upload more than 1.8 billion images every day and we consume more than 4,146,600 videos on YouTube everyday. This abundantly available visual content demands analysing and understanding of them. Computer vision helps in doing that by teaching machines to "see" these images and videos. Apart from automating a lot of functions, computer vision also ensures moderation and monitoring of online visual content. The content available on the internet may fall in any one of these categories: text, visual, and audio, Computer vision uses algorithms especially deep learning algorithms to read and index images. Popular search engines like Google and YouTube use computer vision to scan through images and videos to approve them for featuring. This paper surveys these kinds of deep learning algorithms for Computer Vision tasks such as Classification, Segmentation and Object Recognition.

2. Classification

Steps to perform Image Classification

i. ***Image Pre-processing****:* Restraining unwanted distortions and enhancement the features of an image can improve computer vision[1] models. The image preprocessing steps are: Reading image, Resizing image, and Data Augmentation (ie. Gray scaling of image, Reflection, Gaussian Blurring, Histogram, Equalization, Rotation, and Translation).

ii. ***Detection of an object****:* The Segmentation (localization of an object) of an image and identifying the position of the object of interest is called as 'Detection of an object [2]'.

iii. ***Extracting Feature and training****:* The statistical or deep learning[3] methods are used to identify the most interesting patterns of the image, which might be unique to a particular class, is known as "Feature Extraction" and that will, later on, help the model to differentiate between different classes. This process where the model learns the features from the dataset is called model training.

iv. ***Classification****:* Classification categorizes detected objects into predefined classes by using a suitable

Paper ID : ICESA029 Date of Conference: 12.11.2021 &13.11.2021
International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology
Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

classification technique that compares the image patterns with the target patterns.

Supervised Classification

Selecting sample pixels in an image which are representative of specific classes and then direct the image processing software to use these representative pixels as references for the classification of all other pixels in the image is known as '**Supervised Classification**'. The Representative pixels forms the Training sites (also known as testing sets or input classes) which are selected based on the knowledge of the user. The user also sets the bounds for how similar other pixels must be to group them together. This type of classification uses classification algorithms and regression techniques to develop predictive models. The algorithms contain linear regression, logistic regression, neural networks, decision tree, support vector machine, random forest, naive Bayes, and k-nearest neighbor.

Unsupervised Classification

In Unsupervised Classification, the outputs are based on analysis of an image, without the user providing sample classes. The user can specify which algorithm the software will use and the desired number of output classes, but the user does not help in the classification process, but they must have knowledge of the area being classified Most common algorithms which are used in unsupervised learning[4] are cluster analysis, anomaly detection, neural networks.

Artificial Neural Networks (ANN) on Classification

ANNs are statistical learning algorithms[5] and are used for a variety of tasks no computer vision and speech recognition.

ANNs are implemented as a system of interconnected processing elements, called nodes, which are functionally similar to biological neurons. The interconnections between different nodes have numerical values, called weights, and by altering these values in a systematic way, the network is able to approximate the desired function. A neural network is made up of vertically stacked components called Layers. First layer is input layer. This layer will accept the data. The second layer is called the hidden layer. An ANN may have one or more hidden layers. The last layer is the output layer. The output layer contains the result or the output of the problem. In ANN, a hidden layer is located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through an activation function as the output.

The different kinds of ANNs are convolutional neural network[6], feedforward neural network, probabilistic neural network, time delay neural network, deep stacking network, radial basis function network, and recurrent neural network.

The Convolutional neural network & Recurrent neural network algorithms influences more on Classification of images.

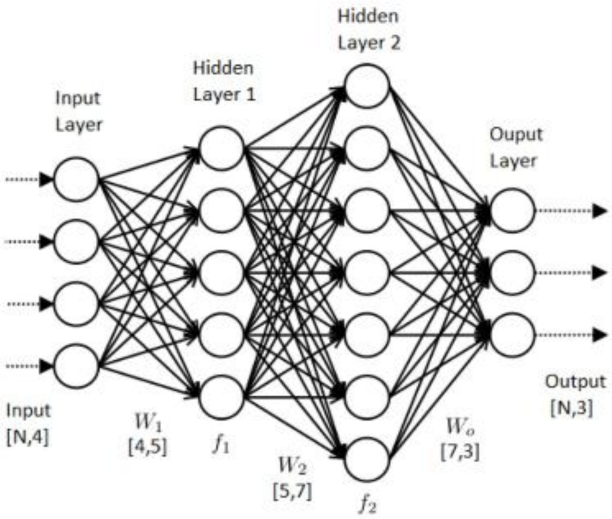


Figure 1

Convolutional Neural Networks (CNN or ConvNet)

CNN or ConvNet is a special type of multi-layer neural network, designed to recognize visual patterns directly from pixel images with minimal pre-processing. CNN is comprised of two elements, namely convolutional layers and pooling layers. There are near-infinite ways to arrange these layers for a given computer vision problem. The elements of a CNN, such as convolutional and pooling layers, are relatively simple to understand. The complex part of using CNN in practice is how to design model architecture that best use these simple elements. The reason why CNN is popular

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022 ISSN: 2348-6600
Reference ID: IJCS-389 PAGE NO: 2657-2667

Paper ID : ICESA029 *Date of Conference: 12.11.2021 &13.11.2021*

International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology

Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

because of its architecture, the best thing is there is no need of feature extraction. The system learns to do feature extraction and the core concept is, it uses convolution of image and filters to generate invariant features which are passed on to the next layer. The features in next layer are convoluted with different filters to generate more invariant and abstract features and the process continues till it gets final feature/output. Commonly used architectures of CNN are AlexNet, LeNet, GoogLeNet, VGGNet, ResNet, and ZFNet,

Support Vector Machine (SVM)

SVM is a powerful and flexible supervised machine learning algorithm. It is extremely popular because of its ability to handle multiple continuous and categorical variables. SVM model is basically a representation of different classes in a hyper-plane in multidimensional space. The hyper-plane will be generated in an iterative manner by the SVM to minimize the error. The intention is to divide the datasets into classes to find a maximum marginal hyper-plane. SVM creates a hyper-plane or a set of hyper-planes in a high dimensional space and good separation between the two classes is obtained by the hyper-plane that has the largest distance to the nearest training data point of any class. The real power of this algorithm depends on the kernel function being used. Regularly used kernels are linear kernel, gaussian kernel, and polynomial kernel.

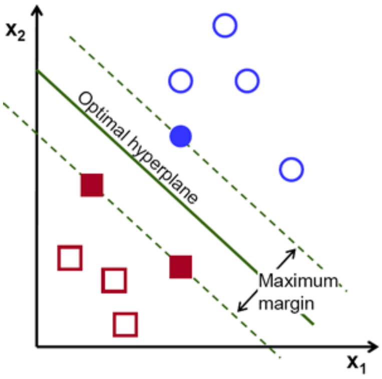


Figure 2

K-Nearest Neighbor Algorithm

The K-Nearest Neighbor is a non-parametric, lazy learning algorithm, where the function is only approximated locally and all computation is deferred until function evaluation. This algorithm simply depends on the distance between feature vectors and it classifies the unknown data points by finding the most common class among the k-closest examples. In order to apply the k-nearest Neighbor classification, we need to define a distance metric or similarity function, like Euclidean distance and Manhattan distance functions. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. **Figure 4** describes the steps in K-Nearest Neighbor Algorithm.

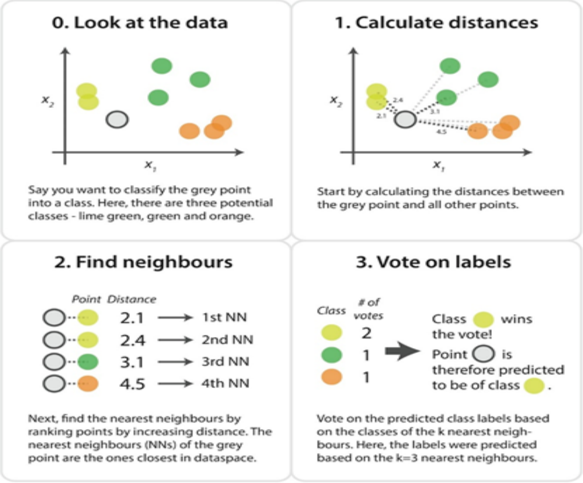

0. Look at the data
x_2
x_1
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.
1. Calculate distances
x_2
2.4
2.1
3.1
4.5
x_1
Start by calculating the distances between the grey point and all other points.
2. Find neighbours
Point Distance
2.1 → 1st NN
2.4 → 2nd NN
3.1 → 3rd NN
4.5 → 4th NN
Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.
3. Vote on labels
Class # of votes
2
1
1
Class wins the vote!
Point is therefore predicted to be of class .
Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Figure 3

Random Forest Algorithm

Random Forest Algorithm is a supervised learning algorithm. It is used for classification[7] and regression. As we know that a forest is made up of trees and more trees creates robust forest, similarly, random forest algorithm creates decision trees on data samples and then gets the

Paper ID : ICESA029 Date of Conference: 12.11.2021 &13.11.2021
International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology
Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

prediction from each of them and finally selects the best solution by means of voting. It is a grouping method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

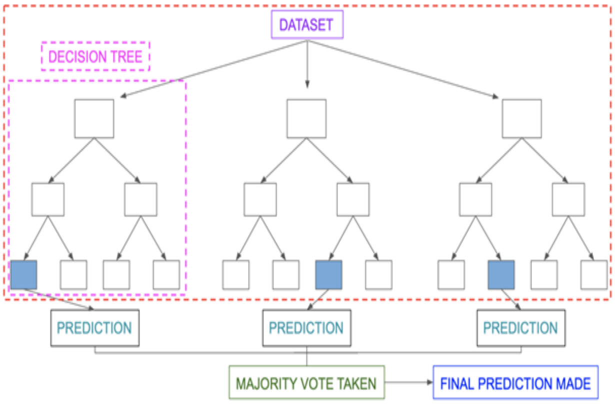


Figure 4

3. Segmentation

Segmentation is the process of dividing an image into different regions based on the characteristics of pixels to identify objects or boundaries to simplify an image and more efficiently analyze it.

Looking at the big picture to understand the scene is one of the complex tasks in Computer Vision. Segmentation plays an important role to understand the scene. Scene understanding is a core computer vision problem is highlighted by the fact that an increasing number of applications nourish from inferring knowledge from imagery. Self-driving vehicles, Human-computer interaction, Virtual reality are some of those applications. With the popularity of deep learning in recent years, many semantic segmentation problems are being dealt with using deep architectures. The CNN beats other approaches in terms of accuracy and efficiency.

The origin of the object could be located at classification, which consists of making a prediction for a whole input image. The next step is localization / detection, which provide not only the classes but also additional information regarding the spatial location of those classes. Finally, **semantic segmentation[8]** achieves fine-grained inference by making dense predictions inferring labels[9] for every pixel, so that each pixel is labeled with the class of its enclosing object region.

The following standard deep networks have made significant contributions to the field of computer vision, as they are often used as the basis of semantic segmentation systems:

CNN	Year	Layers	Accuracy
Toronto's Pioneering deep CNN - **AlexNet**	2012	5 convolutional layers, one max-pooling layer	84.6%.
Oxford's Model - **VGG-16**	2013	a stack of convolution layers with small receptive fields in the first layers instead of few layers with big receptive fields.	92.7%
Google's Network - **GoogLeNet**	2014	22 layers and a newly introduced building block called *inception* module	93.3%.
Microsoft's Model - **ResNet**	2016	152 layers and the introduction of residual blocks.	96.4 %

Table - 1

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022 ISSN: 2348-6600
Reference ID: IJCS-389 PAGE NO: 2657-2667

Paper ID : ICESA029 *Date of Conference: 12.11.2021 &13.11.2021*

International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology

Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

Semantic segmentation

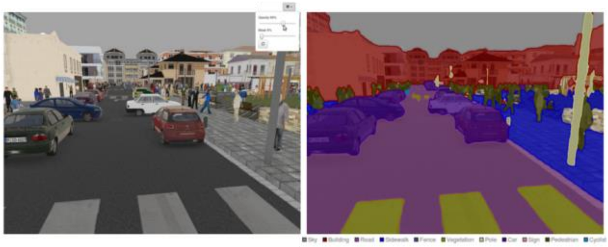
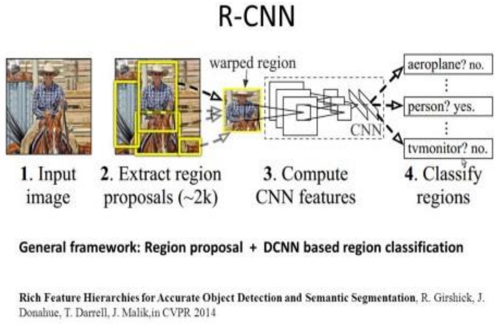


Figure 5

Semantic Segmentation approaches

a. Region-Based Semantic Segmentation

The region-based methods commonly follow the "segmentation using recognition" pipeline, which first extracts free-form regions from an image and describes them, followed by region-based classification. The region-based predictions are transformed to pixel predictions. Highest scoring region is used to label a pixel.



Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, R. Girshick, J. Donahue, T. Darrell, J. Malik,in CVPR 2014

Figure 6

R – CNN (Region based Convolutional Neural Network) performs the semantic segmentation based on the object detection results with region based segmentation. R-CNN first utilizes selective search to extract a large quantity of object proposals and then computes CNN features for each of them. Lastly, it classifies each region using the class-specific linear SVMs. R-CNN can be built on top of any CNN benchmark structures, such as AlexNet, VGG, GoogLeNet, and ResNet.

For the image segmentation task, R-CNN extracted two kinds of features for each region. They are full region feature and foreground region feature and concatenate them for better performance.

b. Fully Convolutional Network (FCN) - Based Semantic Segmentation

FCN learns a mapping from pixels to pixels, without extracting the region proposals. FCN[10] is an extension of the classical CNN. The classical CNN does not allow arbitrary-sized images as input. But FCN have convolutional and pooling layers which allows arbitrary-sized images as input and able to make prediction on them.

FCN Architecture

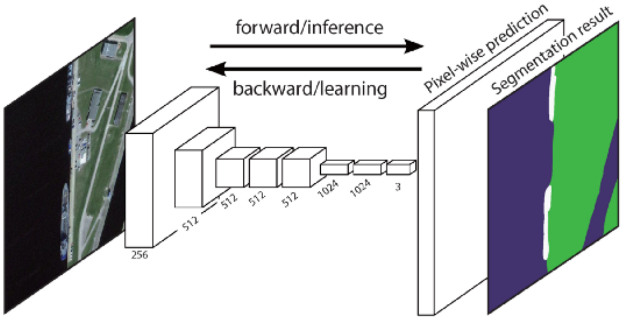


Figure 7

One important concern in this specific FCN is that by propagating through several alternated convolutional and pooling layers, the resolution of the output feature maps is down sampled. So that, the direct predictions of FCN are typically in low resolution. A variety of more advanced FCN-based approaches have been proposed to address this concern, including SegNet, DeepLab-CRF, and Dilated Convolutions.

c. Weakly Supervised Semantic Segmentation

Most of the methods in semantic segmentation based on a large number of images with pixel-wise segmentation masks. However, manually annotating these masks is quite time-consuming, annoying and commercially expensive. Therefore, some weakly supervised methods[11] have recently been proposed, which are dedicated to fulfill the semantic segmentation by utilizing annotated bounding boxes.

Boxsup Training

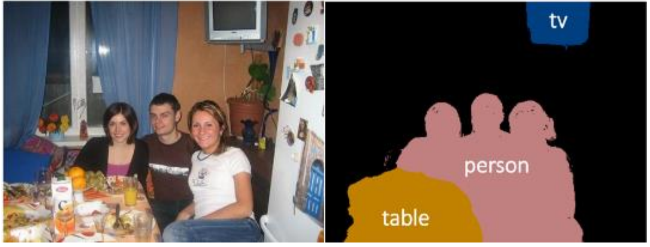


Figure 8

For instance, Boxsup engaged the bounding box annotations as a supervision to train the network and iteratively improve the expected masks for semantic segmentation. It treated the weak supervision limitation as an concern of input label noise and explored recursive training as a de-noising strategy. Pixel-level Labeling interpreted the segmentation task within the multiple-instance learning framework and added an extra layer to restrict the model to assign more weight to important pixels for image-level classification.

4. OBJECT RECOGNITION

A collection of related computer vision tasks that involve identifying objects in digital photographs is known as "Object Recognition" [12]. Mostly it combines the tasks of Image Classification, Segmentation (object localization) and Object detection.

Image classification involves predicting the class of one object in an image. *Object localization[13]* refers to identifying the location of one or more objects in an image and drawing abounding box around their size. *Object detection* combines these two tasks and localizes and classifies one or more objects in an image.

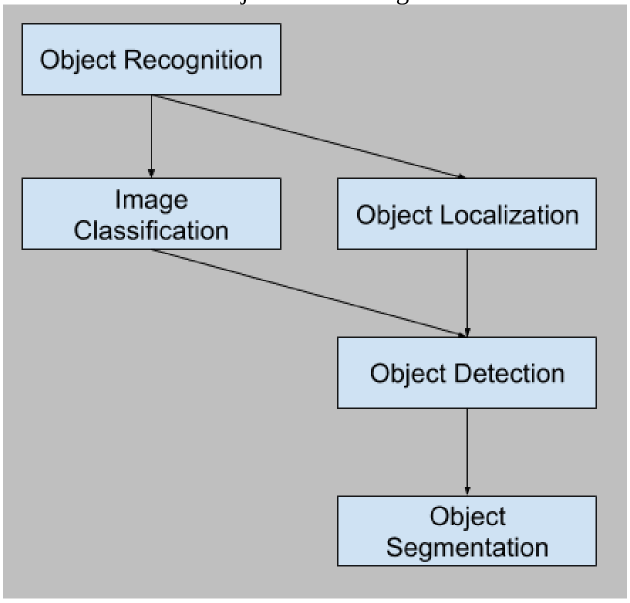


Figure 9

We can differentiate these three computer vision tasks as follows:

Computer Vision Task	Input	Output
Image Classification: Predict the type or class of an object in an image.	An image with a single object, such as a photograph.	A class label (e.g. one or more integers that are mapped to class labels).

IJCS International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
SKRGC Since 2012

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022 ISSN: 2348-6600
Reference ID: IJCS-389 PAGE NO: 2657-2667

Paper ID : ICESA029 Date of Conference: 12.11.2021 &13.11.2021
International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology
Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

Object Localization: Locate the presence of objects in an image and indicate their location with a bounding box.	An image with one or more objects, such as a photograph.	One or more bounding boxes (e.g. defined by a point, width, and height).
Object Detection: Locate the presence of objects with a bounding box and types or classes of the located objects in an image.	An image with one or more objects, such as a photograph.	One or more bounding boxes (e.g. defined by a point, width, and height), and a class label for each bounding box.

Table 2

On this present era R-CNN Model family of Object Recognition algorithms[14] play a vital role in computer vision from 2012. Now a day, YOLO Model Family of algorithms also influences the CV. First we review the R-CNN Model family and then the YOLO family of object recognition algorithms.

R-CNN (*Regions with CNN Features*) Model Family

The R-CNN family includes the techniques R-CNN, Fast R-CNN, and Faster-RCNN designed for object localization and object recognition. The R-CNN was developed by Ross Girshick, et al.

R-CNN

The R-CNN was described in the 2014 paper by Ross Girshick, et al. from UC Berkeley titled “Rich feature hierarchies for accurate object detection and semantic segmentation.”

It may have been one of the first big and successful applications of CNN to the problem of object localization, detection, and segmentation. This approach was demonstrated on benchmark datasets, achieving then state-of-the-art results on the VOC-2012 dataset and the 200-class ILSVRC-2013 object detection dataset.

R-CNN model is comprised of three modules: Region Proposal, Feature Extractor and Classifier.

Region Proposal: Generate and extract category independent region proposals. e.g. candidate bounding boxes.

Feature Extractor. Extract feature from each candidate region. e.g. using a deep convolutional neural network.

Classifier: Classify features as one of the known classes. e.g. linear SVM classifier model.

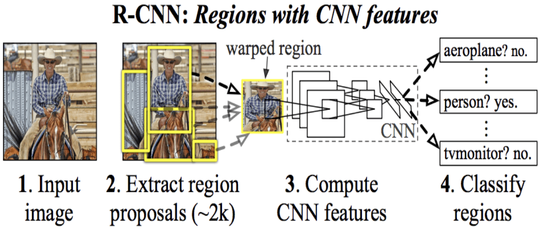


Figure 10

R-CNN proposes candidate regions or bounding boxes of potential objects in the image, which is called “*selective search*”. For feature extractor R – CNN uses the AlexNet deep CNN[15] which won the ILSCRC-2012 image classification competition. The output from this R –CNN was a 4,096 element vector that describes the contents of the image that is fed to a linear SVM for classification; specifically one SVM is trained for each known class. It is a simple and straightforward approach of CNNs. Disadvantage of this approach is that it is slow, requiring a CNN-based feature extraction pass on each of the candidate regions generated by the region proposal algorithm.

Fast R-CNN

In 2015, Ross Girshick, at Microsoft Research, reviewed the limitations of R-CNN, in his paper titled “Fast R-CNN”[16] which can be summarized as follows:

“Training is a multi-stage pipeline - Involves the preparation and operation of three separate models.

Training is expensive in space and time - Training a deep CNN on so many region proposals per image is very slow.

Object detection is slow - Make predictions using a deep CNN on so many region proposals is very slow".

To remove the above limitations Ross Girshick introduced "Fast R-CNN", which is proposed as a single model instead of a pipeline to learn and output regions and classifications directly. Fast R-CNN takes the photograph, a set of region proposals as input that is passed through a deep convolutional neural network. For feature extraction, a pre-trained CNN, such as VGG-16, is used. The end of the deep CNN is a custom layer called a Region of Interest Pooling Layer, or RoI Pooling, that extracts features specific for a given input candidate region. The output of the CNN is then interpreted by a fully connected layer then the model bifurcates into two outputs, one for the class prediction via a softmax layer, and another with a linear output for the bounding box. This process is then repeated multiple times for each region of interest in a given image. The model is significantly faster than R- CNN.

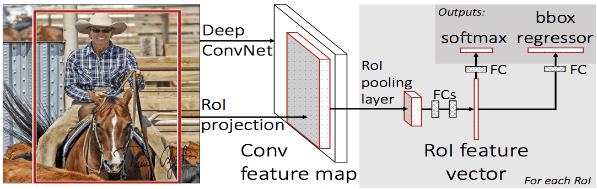


Figure 11

Faster R-CNN

In the 2016, Shaoqing Ren, et al. at Microsoft Research, further improved the Fast CNN for speed of training and detection in the name "**Faster R-CNN [17]".** The Faster R-CNN architecture was designed to propose and refine region proposals as part of the training process, referred to as a Region Proposal Network[18](RPN). These regions are then used in concert with a Fast R-CNN model in a single model design. These improvements both reduce the number of region proposals and speed up the test-time operation of the model to near real-time with then state-of-the-art performance.

The Faster R-CNN architecture is comprised of two modules:

Region Proposal Network: CNN for proposing regions and the type of object to consider in the region.

Fast R-CNN: CNN for extracting features from the proposed regions and outputting the bounding box and class labels. Both modules operate on the same output of a deep CNN. The RPN acts as an attention method for the Fast R-CNN, informing the second network of where to look or pay attention. This Faster R-CNN architecture is the peak of the family of models and continues to achieve near state-of-the-art results on object recognition tasks.

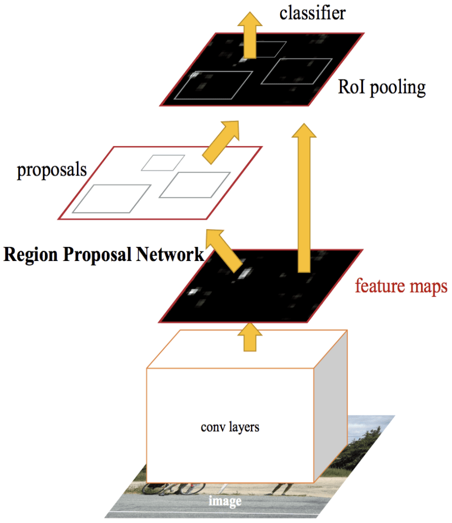


Figure 12

YOLO Model Family

Another popular family of object recognition models is referred to as YOLO or "*You Only Look Once*", developed by Joseph Redmon, et al in the year 2015. The R-CNN models may be more accurate, yet the YOLO family of

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022 ISSN: 2348-6600
Reference ID: IJCS-389 PAGE NO: 2657-2667

Paper ID : ICESA029 *Date of Conference: 12.11.2021 &13.11.2021*

International Conference on Emerging Vistas of Soft Computing and Advanced Communication Technology

Department of Computer Science & BCA, Mangayarkarasi College of Arts and Science for Women, Madurai, Tamil Nadu, India & Association with ICT Academy, Chennai, Tamil Nadu, India.

models is fast, much faster than R-CNN, achieving object detection in real-time.

YOLO

The YOLO[19] involves a single neural network trained end to end that takes a photograph as input and predicts bounding boxes and class labels for each bounding box directly. This technique offers lower predictive accuracy, although operates at 45 frames per second and up to 155 frames per second for a speed-optimized version of the model.

The model works by first dividing the input image into a grid of cells, where each cell is responsible for predicting a bounding box. If the center of a bounding box falls within the cell, each grid cell predicts a bounding box involving the x, y coordinate and the width and height and the confidence. A class prediction is also based on each cell.

For instance, an image may be divided into a 7×7 grid and each cell in the grid may predict 2 bounding boxes, resulting in 94 proposed bounding box predictions. The class probabilities map and the bounding boxes with confidences are then combined into a final set of bounding boxes and class labels.

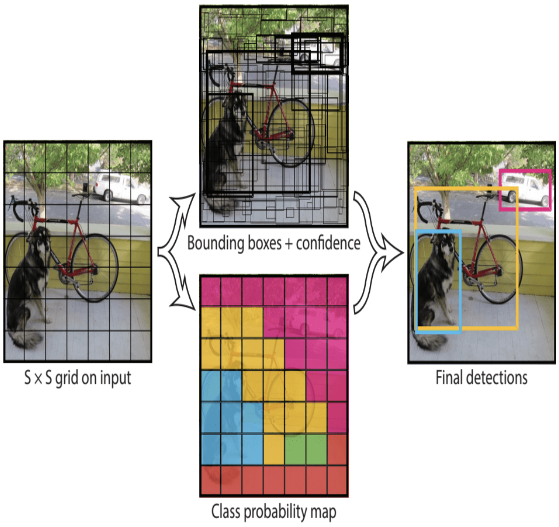


Figure 13

YOLOv2 and YOLOv3

The YOLO model was updated by Joseph Redmon and Ali Farhadi in an effort to further improve model performance in their 2016 paper titled "YOLO9000: Better, Faster, Stronger". This variation of the model is referred to as YOLO v2[20], an instance of the model is described that was trained on two object recognition datasets in parallel, capable of predicting 9,000 object classes, hence given the name "*YOLO9000*."

Similar to Faster R-CNN, YOLOv2 model makes use of anchor boxes, pre-defined bounding boxes with useful shapes and sizes that are tailored during training. The preference of bounding boxes for the image is pre-processed using a k-means analysis on the training dataset. Significantly, the predicted representation of the bounding boxes is changed to allow small changes to have a less vivid effect on the predictions, resulting in a more stable model. Rather than predicting position and size directly, offsets are predicted for moving and reshaping the pre-defined anchor boxes relative to a grid cell and dampened by a logistic function.

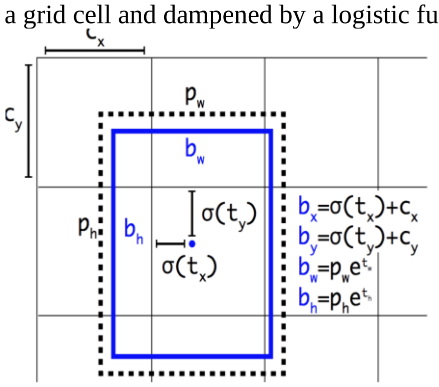


Figure 14

Further improvements to the model were proposed by Joseph Redmon and Ali Farhadi in their 2018 paper titled "YOLO v3[21] An Incremental Improvement." The improvements were reasonably minor, including a deeper feature detector network and minor representational changes.

5. DISCUSSION

Challenges we face in Computer Vision

Issue on Reasoning: Recent neural network-based algorithms are complex system whose functions are often difficult to understand, it becomes tough to find the logic behind any task. This lack of reasoning creates a real challenge for computer vision experts who try to define any feature in an image or video.

Ethics and Privacy: Vision powered surveillance is a serious threat to privacy in a lot of countries. It exposes people to unauthorized use of data. Face recognition and detection is prohibited in some countries because of these problems.

Fake Content: Computer vision in the wrong hands can lead to dangerous problems. Anybody with access to powerful data centre is capable of creating fake images, videos or text content.

Attacks: When an attacker creates a faulty machine learning model, they intend the machine using it to fail. These faulty models are difficult to identify and can cause serious damage to any system.

Once computer vision specialists can resolve the current problems of the domain, we can expect a truthful system that automates content moderation and monitoring. With communal giants like Facebook, Google, Apple and Microsoft investing in computer vision, it's only a matter of time before it takes over the international market.

References

[1] R. Girshick, "Fast R-CNN," in *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV '15)*, pp. 1440–1448, December 2015.

[2] T. Chen, S. Lu, and J. Fan, "S-CNN: Subcategory-aware convolutional networks for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[3] Jason Brownlee, "Step-by-step tutorials on deep learning neural networks for computer vision in python with Keras", Google Books, 4 April 2019.

[4] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Communications of the ACM*, vol. 54, no. 10, pp. 95–103, 2011.

[5] N. Doulamis, "Adaptable deep learning structures for object labeling/tracking under dynamic visual environments," *Multimedia Tools and Applications*, pp. 1–39, 2017.

[6] N. Doulamis and A. Voulodimos, "FAST-MDL: Fast Adaptive Supervised Training of multi-layered deep learning models for consistent object tracking and classification," in *Proceedings of the 2016 IEEE International Conference on Imaging Systems and Techniques, IST 2016*, pp. 318–323, October 2016.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 580–587, Columbus, Ohio, USA, June 2014.

[8] X. Wu, R. He, Z. Sun, and T. Tan, A light CNN for deep face representation with noisy labels, https://arxiv.org/abs/1511.02683.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 3431–3440, IEEE, Boston, Mass, USA, June 2015.

[10] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. V. Gool, "Weakly Supervised Cascaded Convolutional Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5131–5139, Honolulu, HI, July 2017.

[11] Jason Brownlee,"A Gentle Introduction to Object Recognition With Deep Learning", on May 22, 2019 in Deep Learning for Computer Vision

[12] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - Weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 685–694, June 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *Computer Vision – ECCV 2014*, vol. 8691 of *Lecture Notes in Computer Science*, pp. 346–361, Springer International Publishing, Cham, 2014.

[14] V. Nair and G. E. Hinton, "3D object recognition with deep belief nets," in *Proceedings of the NIPS*, 2009.

[15] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 1–9, Boston, Mass, USA, June 2015.

[16] R. Girshick, "Fast R-CNN," in *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV '15)*, pp. 1440–1448, December 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[19] Wei Fang (Member IEEE), Wang & Peiming "Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments", IEEE Access, Dec 24,2019.

[20] Juain Du,"Understanding of Object Detection Based on CNN family and YOLO", IOP conference series 1004(2018)012029.

[21] Wei Fang (Member IEEE), Wang & Peiming "Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments", IEEE Access, Jan 6,2020.