INTERNATIONAL JOURNAL IN ENGINEERING SCIENCES

Journal Homepage: https://www.ijesjournal.com

OPEN ACCESS

Research Article

Object Detection Using Yolo And Tensor Flow

M. Chinnarao, R. Goutham Sai Kalyan, T. Naga Pravallika, B. Srinivas

Computer Science And Engineering, Lingayas Institute of Management and Technology

ARTICLE INFO

Published: 16 June 2024

Keywords:
Deep Learning, Neural Networks, Object Detection, YOLOv3, Residual Networks.

DOI:
10.5281/zenodo.11825059

ABSTRACT

Object detection methods aim to identify all target objects in the target image and determine the categories and position information in order to achieve machine vision understanding. Numerous approaches have been proposed to solve this problem, mainly inspired by methods of computer vision and deep learning. However, existing approaches always perform poorly for the detection of small, dense objects, and even fail to detect objects with random geometric transformations. In this study, we compare and analyse mainstream object detection algorithms and propose a multi-scaled deformable convolutional object detection network to deal with the challenges faced by current methods. Our analysis demonstrates a strong performance on par, or even better, than state of the art methods. We use deep convolutional networks to obtain multi-scaled features, and add deformable convolutional structures to overcome geometric transformations. We then fuse the multi-scaled features by up sampling, in order to implement the final object recognition and region regress. Experiments prove that our suggested framework improves the accuracy of detecting small target objects with geometric deformation, showing significant improvements in the trade-of between accuracy and speed.

INTRODUCTION

The main purpose of object detection is to identify and locate one or more effective targets from still image or video data. It comprehensively includes a variety of important techniques, such as image processing, pattern recognition, artificial intelligence and machine learning. It has broad application prospects in such areas such as road traffic accident prevention, warnings of dangerous goods in factories, military restricted area monitoring and advanced human–computer interaction. Since the application scenarios of multi-target detection in the real world are usually complex and variable, balancing the relationship between accuracy and computing costs is a difficult task.

The object detection process is traditionally established by manually extracting feature models, where the common features are represented by HOG (histogram of oriented gradient), SIFT

***Corresponding Author:** M. Chinnarao
Address: *Computer Science And Engineering, Lingayas Institute of Management and Technology*
Email : reddygoutamsai@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

(scale-invariant feature transform), Haar (Haar-like features) and other classic algorithms based on grayscale. Following feature extraction, the SVM (support vector machine) or Adaboost algorithms are used for classification in order to obtain target information. These traditional extracting feature models are only able to determine low-level feature information, such as contour information and texture information, and have limitations in detecting multiple targets under complex scenes due to their poor generalization performance. However, object detection models, such as the R-CNN (region-based convolutional neural networks) series and the YOLO (you only look once) or SSD (single shot multiBox detection) models based on the deep learning CNN (convolutional neural network) features are more well-known.

Deep learning CNN models not only extract the detail texture features from pre-level convolution networks, but are also able to obtain higher-level information from the post-level convolution layer. Following the traditional CNN process, the R-CNN series uses an enumeration method to presuppose the target candidate region in the feature map, gradually fine-tuning the position information and optimizing the object position for classification and recognition. In contrast, other object detection models will simultaneously predict the bounding box and classification directly in the feature map by applying different convolution sets. the R-CNN model has two operation stages (candidate region proposal and further detection) that allow for higher detection accuracy, while SSD and YOLO are able to directly detect the classification and position information, improving the detection speed. We propose a novel multi-scaled deformable convolution network model to deal with the trade-of between accuracy and speed in object detection. The multi-scaled deformable convolutional neural network uses a new convolution method that has two offsets for image feature generation that are more sensitive to object deformation information. Additionally, the ability to detect objects that have geometrical deformations is improved. Secondly, feature fusion operations are performed on the multiple scale feature maps in the final detection. The image information of different scaled feature maps is simultaneously used to predict the classification and position information. Tis modification ensures the detection speed, enhances the target information of small objects, and also improves the accuracy of object detection.

The key contributions of our work are as follows:

1. The novel deformable convolution structure replaces the ordinary normal convolution operation for object detection. It effectively lets the CNN improve the generalization ability of extracting image features under different geometric deformations. Also, the new network automatically trains the offset of the convolution without wasting a large amount of computation time and cache space. Tus, significant performance gains on computer vision tasks, such as object detection and semantic segmentation, are observed.

2. An up-sample is applied to the feature pyramid to merge the multi-scaled feature information. Tis increases the accuracy of small target object detection by avoiding the loss of information of small target objects after multiple convolution and pooling operations. It also provides an important scheme for the detection of dense objects with overlapping occlusion in complex scenes.

MOTIVATION

The intension of object detection is to develop computational models that provide the most fundamental information needed by computer vision applications.

- The motive of object detection is to recognize and locate (localize) all known objects in a scene.

- Preferably in 3D space, recovering pose of objects in 3D is very important for robotic control systems.
- The information from the object detector can be used for obstacle avoidance and other interactions with the environment.
- One of the best examples of why you need object detection is the high-level algorithm for autonomous driving:
- In order for a car to decide what to do next: accelerate, apply brakes or turn, it needs to know where all the objects are around the car and what those objects are that requires object detection.
- You would essentially train the car to detect known set of objects: cars, pedestrians, traffic lights, road signs, bicycles, motorcycles, etc.
- While police or any other officials while watching the CCTV footage they want to know the objects that are present there.

So, in this kind of situations also we can use object detection.

OBJECT DETECTION

- **Object detection** is merely to recognize the object with bounding box in the image, where in image classification, we can simply categorize(classify) that is an object in the image or not in terms of the likelihood (Probability).
- **Object detection** is considered one of the noteworthy areas in the deep learning and Computer vision. Object detection has been determined the numerous applications in computer vision such as object tracking, retrieval, video surveillance, image captioning, Image segmentation, Medical Imagine and several greater number other applications as well.

DEFORMABLE CONVOLUTION

Deformable convolution introduces the mechanism of deformable module, which has learn-able shape to adapt to the changes of features. Conventionally, the shapes of kernels and samples in convolution, defined by the sampling matrix, is fixed from the start.

COMPUTER VISION

Computer Vision (CV) is a field of Artificial Intelligence (AI) that deals with computational methods to help computers understand and interpret the content of digital images and videos. Hence, computer vision aims to make computers see and understand visual data input from cameras or sensors.

YOLO

YOLO is an algorithm that uses neural networks to provide real-time object detection. This algorithm is popular because of its speed and accuracy. It has been used in various applications to detect traffic signals, people, parking meters, and animals.

METHODOLOGIES

ALGORITHM

YOLO algorithm is used here.

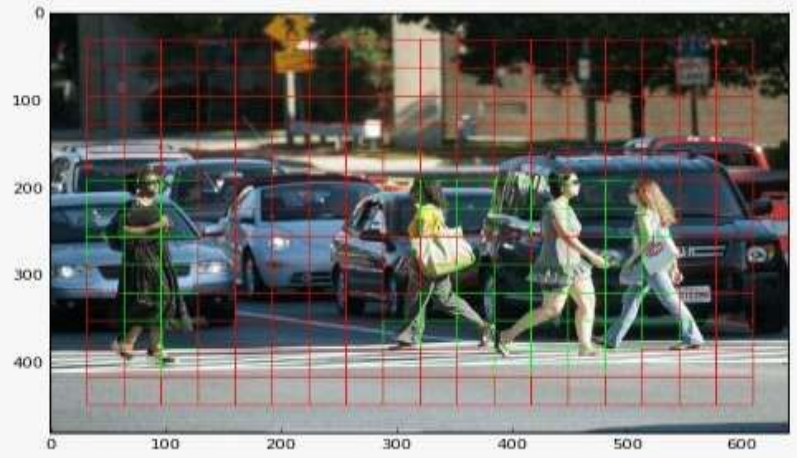
YOLO means You Only Look Once.

YOLO algorithm works using the following three techniques:

1. Residual blocks
2. Bounding box regression
3. Intersection Over Union (IOU)

1. Residual blocks

First, the image is divided into various grids. Each grid has a dimension of S x S. The following image shows how an input image is divided into grids.



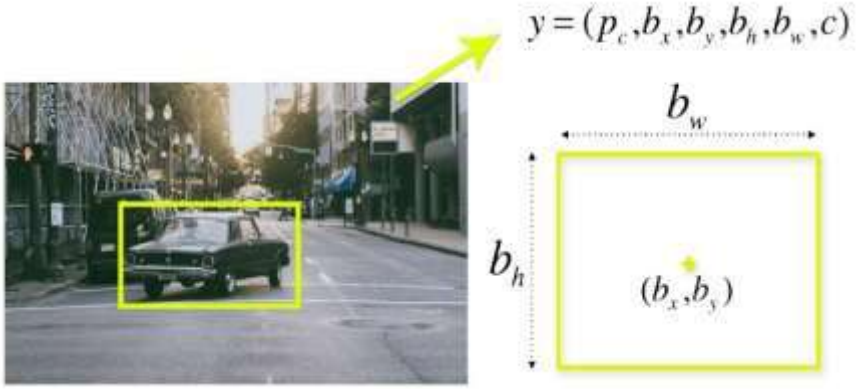
Residual Blocks

2. Bounding box regression

- A bounding box is an outline that highlights an object in an image.
- Every bounding box in the image consists of the following attributes:

  Width (bw)

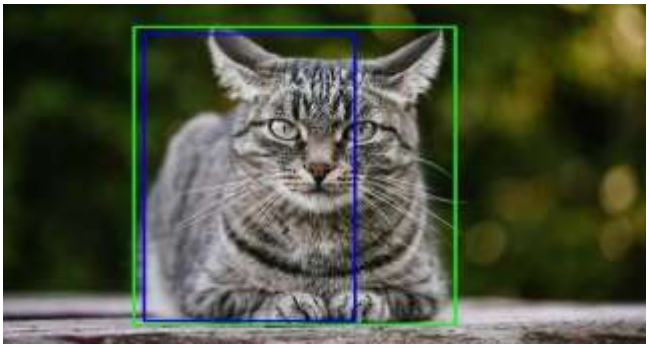
  Height (bh)
- Bounding box center (bx,by)
- The following image shows an example of a bounding box. The bounding box has been represented by a yellow outline.



Bounding Box Regression

3. Intersection over union (IOU)

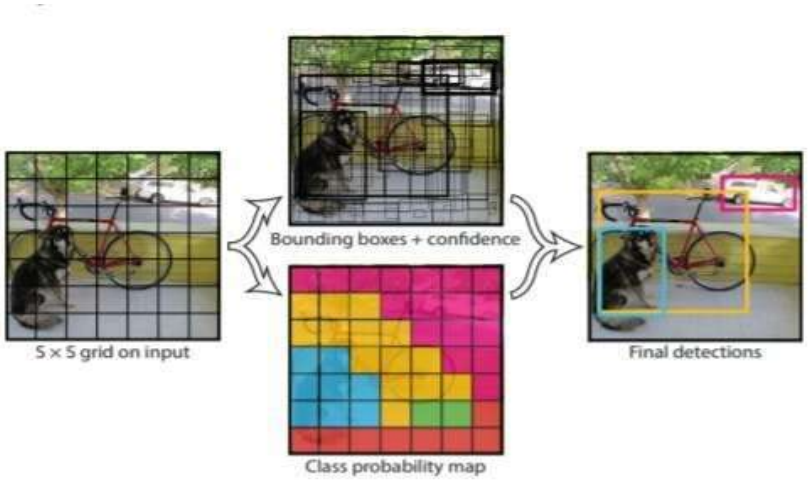
- Intersection over union (IOU) is a phenomenon in object detection that describes how boxes overlap.
- YOLO uses IOU to provide an output box that surrounds the objects perfectly.
- Each grid cell is responsible for predicting the bounding boxes and their confidence scores.
- The IOU is equal to 1 if the predicted bounding box is the same as the real box. This mechanism eliminates bounding boxes that are not equal to the real box. The following image provides a simple example of how IOU works.



Intersection over union (IOU)

- Combination of the three techniques:

The following image shows how the three techniques are applied to produce the final detection results.



Our image object detector adopts YOLO's backbone network and adds the new trick in convolution operation and feature information fusion. The overall framework is shown as Fig. 1. The first backbone network is the Darknet53 network. As a new network for performing feature extraction, it is a hybrid approach combining the network used in YOLO v2, Darknet-19, and the newer residual network tactics. The network, which is larger, uses successive 3×3 and 1×1 convolutional layers, with shortcut connections. In addition, we add three deformable convolution layers before the convolutional layers with a size of 52×52, 26×26 and 13×13 to modify the feature extraction (see the yellow section in Fig. 1). The second element is the detection network section. The YOLO detection network divides the input image into 7×7 grids. If the centre position of the ground truth falls within a certain grid, three bounding boxes and their confidences, as well as 20 class probabilities, are predicted for each grid. We also use the convolutional set, which is made up of 3×3 and 1×1 convolutional layers, in order to control the output, which includes 20 types of classification information, three frames positions and the IOU position. The new trick mentioned above refers to the detection network performing the above operations on three different feature map scales (13×13, 26×26 and 52×26, respectively). The upper-level feature maps will be up-sampled and merged with the low-level layer features by the channel (see the red section in Fig. 1)

Deformable convolutional network

For object recognition of real scenes, an inevitable challenge arises from the recognition errors caused by changes in the shape, size, and posture of objects caused by the motion or different observation angles. Generally speaking, there are two common methods dealing with this question. The first is the data argument, which artificially changing the size, shape and rotation angle of the object in advance, and enhancing the diversity of the data in order to simulate the deformation of the object. However, this method increases the cost of data pre-processing, and the data will never cover all real application scenarios. Thus, the generalization ability of the model will be reduced to some extent. Another method is the use of a transform invariant feature algorithm, such as SIFT. Yet this handcrafted design of invariant features and algorithms can be difficult, or perhaps infeasible, for overly complex transformations, even when they are known. To solve the above problems, this study proposes the idea of applying a deformable convolution network to the one-step object detection network, and changes the fixed geometry of the convolution kernel in the traditional convolutional network, in order to enhance the modelling ability for the geometric transformation of detected objects.

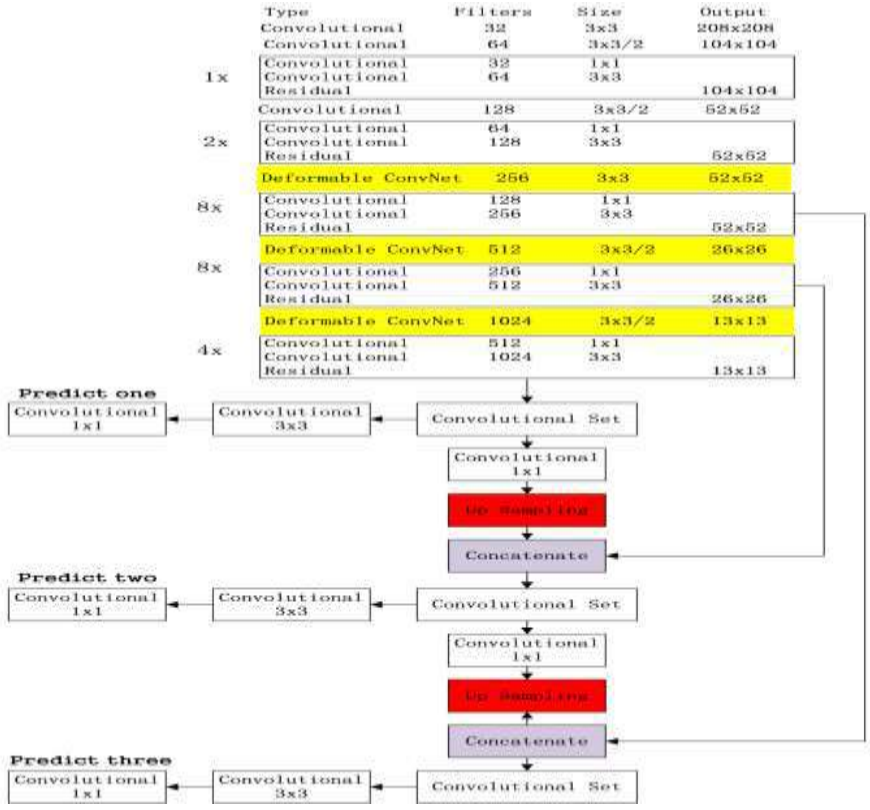


Figure 1

Architecture of the multi-scaled deformable convolutional neural network framework. The framework mainly consists of two components. (1) The backbone network based on Darknet53, which includes residual structures and deformable convolution. (2) The object detection network, based on multi-scaled detection and feature fusion. A common convolution operation performs sampling in the input feature map X with a regular grid R, and sums the sample values under the weights, w. The grid R defines the size and expansion of the receptive field. For example, a 3×3 convolution kernel with an expansion size of 1 can be defined as follows:

$$R = \{(-1, -1), (-1,0) \ldots. (0,1), (1,1)\}, \quad (1)$$

For every output y(Po), the sampling must be performed with nine positions from X. These nine positions are in the shape of a grid diffused around a centre position X(Po). The coordinates $(-1, -1)$ represent the upper left corner of X(Po), while (1,1) represent the lower right corner, with the remaining follow the same representation. Under traditional convolution, for each position Po on the output feature map Y, we output the feature map formula Y as:

$$Y(Po)= \Sigma w(Pn).X(P0+Pn) \quad (2)$$

Under the deformation convolution, for each output y(Po), nine positions are sampled from X, in the shape of the grid that is diffused around centre position X(Po). Subsequently, a new parameter is added. The parameter ΔPn allows the points of sampling to be diffused into an irregular shape as follows:

$$Y(P0) =\Sigma w(Pn)X(P0 + Pn + \Delta pn) \quad (3)$$

The new sampling is located at an irregular position with offset $Pn+ \Delta Pn$. The offset ΔPn is usually a decimal, while the X(Po) on the feature map is always a whole number. Formula (3) can be realized by bilinear interpolation:

$$X(P) = \Sigma G(q,p).X(q) \quad (4)$$

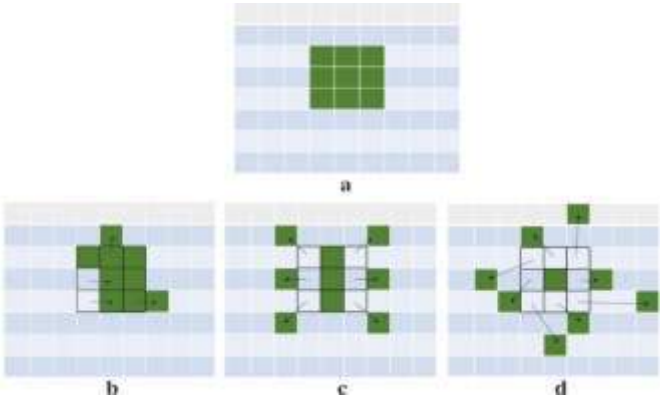
where P represents an arbitrary (decimal) position ($P = Po+ Pn+ \Delta Pn$ in Eq. (2)),

q enumerates all global spatial positions in the feature map X, and G is a bilinear interpolated

kernel. Note that, G is two-dimensional and is divided into two one-dimensional cores:

$$G(p,q) = g(qx,qy).g(qy,py) \quad (5)$$

where g(a,b)=max(0,1−|a−b|) Formula (3) can be calculated quickly, as G(q,p) is non-zero only for some q. Finally, the calculation positions of the convolution kernel on the image will be changed from the original 3 × 3 squared position, as shown in Fig. 2a.



Different calculation positions under traditional convolution or deformable convolution. a Original calculation position of convolution. b–d Different convolution positions changes after deformable convolutional networks are applied.

The fusion of multi scaled features

For object detection of real scenes, the accurate detection of small target objects will determine whether detection information is lost. Although a sampling operation based on convolutional networks already includes robustness to the changes of object size, it is often not sensitive enough for finer-grained small object detection. When object detection is performed on the feature map based on CNN, the feature information of the lower layer is less abundant, but the position information is more accurate. The semantic information of the upper layer is observed at a greater amount, yet the position information is often given at lower amount, as the feature map becomes smaller after pooling layers. Therefore, using the different sizes of the feature maps in the CNN network to better detect objects under different sizes, particularly small target objects that are prone to miss detection, is important for object detection performance. Many studies have investigated how to overcome this challenge, with the easiest way being the application of a data argument, which changes the image size in many different scales.

This involves resizing the images to different scales and training them in the convolutional network to adapt to different scales of object detection. As larger images

increase memory storage and calculations, most experiments resize the images during the testing stage. Despite this, time and memory consumption cannot be avoided. Another method, similar to SPP (spatial pyramid pooling) net, has also been applied to the R-CNN series. For example, Fast R-CNN and Faster R-CNN use the spatial pyramid pooling layer to pool the regions of images of any size using different pooling grids. They then generate fixed-length feature vectors for further classification and regression.

The most widely used and popular method is similar to the SSD network, which introduces a regional detection mechanism on different scales of feature maps. It is reasonable to say that the detection in all different scales will obtain the most comprehensive scale information. Yet considering that low-level feature maps contain poor semantic information, the corresponding detection calculations will slow down the speed. Thus, the SSD network drops the previous low-level features and begins detection with conv4_3. In addition,

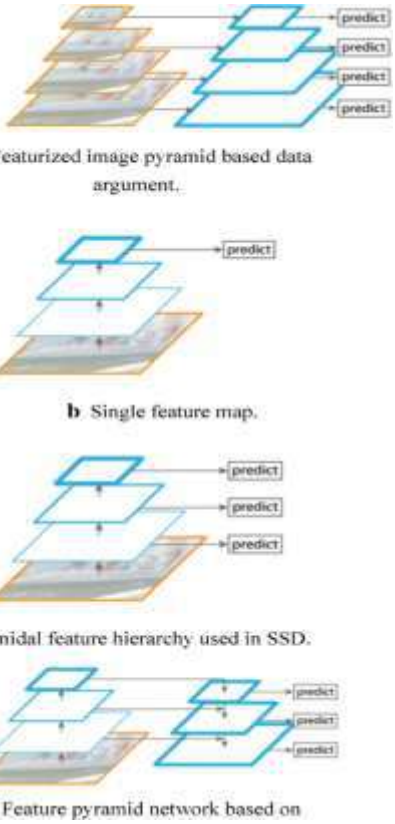
some convolutional layers are added behind conv4_3 to generate additional feature maps for higher-level semantics. However, the results of the final experiment show that this is limited for the detection of small objects. The detection accuracy is poor, and far less than that of YOLO v3.

Also, it is proposed recently that convolution kernels of different sizes could be used to predict classes and bounding boxes of multi-scale objects directly in the last feature map of a deep CNN for rapid object detection with acceptable precision loss is achieved. Based on the research of the FPN network, this study combines the top-level features with the low-level features using upper sampling. In addition, we use the concatenate method instead of the direct addition between feature pixels, and we achieve the fusion between high-level features and low-level features by extending the dimension of the feature map. The new model independently performs the prediction in multiple layers and controls the amount of computation, in order to better utilize the multi-scaled feature map information and further refine the object detection results. Figure 3 presents the research concepts described in this section. The multi-scaled feature fusion adopted in this paper mainly comes from Fig. 3d, and includes the following two steps.

The bottom up path

The feed-forward calculation of the CNN is denoted as the bottom-up path. The feature map is calculated using a convolution kernel, and generally becomes smaller and smaller. This study will take the output of some features as the same as the original size, known as the "Same Network Stage". The above process involves defining a pyramid level for each of the same network phases, and then selecting the output of the last layer of each phase as a reference set for the feature map. In particular, for the residual depth network, we select the activation output of the last residual structure of the "Same Network Stage" as the reference.

These residual module's outputs are denoted as {C3, C4, C5}, corresponding to the outputs of conv3, conv4, and conv5. It is important that their output scales have different pixel sizes of {52×52, 26×26, 13×13}, and the previous pixel size is twice that of the following. Considering the memory usage problem and cross-semantic information in the underlying feature map, conv1 and conv2 are not included in the pyramid.



a Featurized image pyramid based data argument.

b Single feature map.

c Pyramidal feature hierarchy used in SSD.

d Feature pyramid network based on

a Using an image pyramid to build a feature pyramid. **b** Recent detection systems have opted to use only single scale features for faster detection. **c** an alternative is to reuse the pyramidal feature hierarchy computed by ConvNet, as if it were a featurized image pyramid. **d** the feature pyramid network (FPN) is fast, as in b and c, but more accurate due to the fusion and multiple detection under different scaled features. The feature maps are indicated by blue outlines and thicker outlines denote semantically stronger features.

Top down path and horizontal connection

The purpose of this step is to up-sample a more semantic high-level feature map, such that the

features are laterally connected to the features of the previous. Thus, the high level features are enhanced. It is worth noting that the two-layer features from the lateral connection must have an equal spatial size. Tis up-sampling can be performed using, for example, nearest neighbour up-sampling or bilinear interpolation. Once this is carried out, the layer features are combined with the corresponding previous layer feature. Note that the previous layer has to undergo a 1×1 convolution kernel, in order to change the number of channels to that of the last layer in the FPN. This study also uses the convolution feature concatenate method to perform feature fusion after the up-sampling operation. Tis expands and supplements the low-level feature information by increasing the number of channels rather than directly performing the addition between pixels, as in the FPN.

In fact, the concatenate operation is the combination of information channels, which means that the feature dimension of the image itself is increased. We hope to find more information about the location characteristics of the object in the added features. Performing the addition directly between pixels does not change the number of features; it only adds more information to each original dimension. We found that the concatenate operation can skip the process (note that the FPN will use a 1×1 convolution kernel to change the number of channels in order to prepare for addition between feature maps), yet addition will require less computational work in subsequent convolution operations. Finally, testing proves that the concatenate operation causes just a slight improvement compared to addition, increasing the MAP by 0.02.

However, it makes the network structure simpler and easier to understand, and this study thus uses the concatenate operation for the object detection network. For the residual depth network structure, we first use addition for feature fusion by adding a 1×1 convolution kernel to generate the double-channel feature map for the pixel-to-pixel fusion operation with the previous layer in the last of C4 and C5 layers, just as in the original FPN. The specifc network structure is shown in the Fig. 4a. We then aim to use the concatenate between previous layers and the last C4 or C5 layers following the up-sampling operation to realize the feature fusion. The specific network structure is shown in the Fig. 4b. Finally, the merged feature map is processed with a 3×3 convolution kernel to generate the final required feature map (in order to eliminate the aliasing effect of the up sampling). The fused feature layers corresponding to the {C3, C4, C5} layer is {P3, P4, P5}, and the corresponding layer space sizes are the same.

RESULTS

- If we pass an image as an input, then we will get the input as shown in above figure.
- We will get the output image with the objects identified in it.
- In the above figure, we can able to see that the objects are detected and are named with the corresponding object.
- Similarly, we can pass any other image and we will get the corresponding output i.e., with their object's names identified.

CONCLUSION AND FUTURE SCOPE

Based on the trick of deformable convolutional networks, this study proposes a new multi-scaled deformable convolutional object detection network structure. Tis network uses a deformable convolution structure instead of an ordinary

convolution operation in order to increase the learning ability of the model with respect to object geometric deformation, as well as increasing the accuracy of object detection. This study also uses multi-scaled feature maps that combine low-level features by up-sampling to extract target object position information. Tis increases the ability of the model to detect small target objects and dense objects, and also greatly makes up for the defect in missing detections, which is always present in other object detection models.

This object detection is used at

- **Autonomous driving**: YOLO algorithm can be used in autonomous cars to detect objects around cars such as vehicles, people, and parking signals. Object detection in autonomous cars is done to avoid collision since no human driver is controlling the car.
- **Wildlife**: This algorithm is used to detect various types of animals in forests. This type of detection is used by wildlife rangers and journalists to identify animals in videos (both recorded and real-time) and images. Some of the animals that can be detected include giraffes, elephants, and bears.
- **Security**: YOLO can also be used in security systems to enforce security in an area. Let's assume that people have been restricted from passing through a certain area for security reasons. If someone passes through the restricted area, the YOLO algorithm will detect him/her, which will require the security personnel to take further action.

REFERENCES

1. Xia, Z. (2019). An Overview of Deep Learning. Deep Learning in Object Detection and Recognition, 1-18. doi:10.1007/978-981-10-5152-4_1
2. Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. IEEE Transactions on Signal Processing, 39(9), 2101-2104. doi:10.1109/78.134446
3. Real-Time Object Detection for Aiding Visually Impaired using Deep Learning. (2020). International Journal of Engineering and Advanced Technology Regular Issue, 9(4), 1600-1605. doi:10.35940/ijeat.d8374.049420
4. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2016.91
5. Budiharto, W., Gunawan, A. A., Suroso, J. S., Chowanda, A., Patrik, A., & Utama, G. (2018). Fast Object Detection for Quadcopter Drone Using Deep Learning. 2018 3rd International Conference on Computer and Communication Systems (ICCCS). doi:10.1109/ccoms.2018.8463284
6. Li, X., Wang, J., Xu, F., & Song, J. (2019). Improvement of YOLOv3 Algorithm in Workpiece Detection. 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). doi:10.1109/cyber46603.2019.9066490
7. Zhao, L., & Wan, Y. (2019). A New Deep Learning Architecture for Person Detection. 2019 IEEE 5th International Conference on Computer and Communications (ICCC). doi:10.1109/iccc47050.2019.9064172
8. Lu, Z., Lu, J., Ge, Q., & Zhan, T. (2019). Multi-object Detection Method based on YOLO and ResNet Hybrid Networks. 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM). doi:10.1109/icarm.2019.8833671
9. Zhang, D. (2018). Vehicle target detection methods based on color fusion deformable part model. EURASIP Journal on Wireless Communications and Networking, 2018(1). doi:10.1186/s13638-018-1111-8

10. Kang, M., Leng, X., Lin, Z., & Ji, K. (2017). A modified faster R-CNN based on CFAR algorithm for SAR ship detection. 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP). doi:10.1109/rsip.2017.7958815
11. Hung, P. D., & Kien, N. N. (2019). SSD-Mobilenet Implementation for Classifying Fish Species. Advances in Intelligent Systems and Computing Intelligent Computing and Optimization, 399-408. doi:10.1007/978-3-030-33585-4_40
12. Castiblanco, C., Rodriguez, J., Mondragon, I., Parra, C., & Colorado, J. (2014). Air Drones for Explosive Landmines Detection. ROBOT2013: First Iberian Robotics Conference Advances in Intelligent Systems and Computing, 107-114. doi:10.1007/978-3-319-03653-3_9
13. Ma, H., Liu, Y., Ren, Y., & Yu, J. (2019). Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3. Remote Sensing, 12(1), 44. doi:10.3390/rs12010044
14. Khong, L. M., Gale, T. J., Jiang, D., Olivier, J. C., & Ortiz-Catalan, M. (2013). Multi-layer perceptron training algorithms for pattern recognition of myoelectric signals. The 6th 2013 Biomedical Engineering International Conference. doi:10.1109/bmeicon.2013.6687665
15. Andre, T., Neuhold, D., & Bettstetter, C. (2014). Coordinated multi-robot exploration: Out of the box packages for ROS. 2014 IEEE Globecom Workshops (GC Wkshps). doi:10.1109/glocomw.2014.7063639

HOW TO CITE: M. Chinnarao, R. Goutham Sai Kalyan, T. Naga Pravallika, B. Srinivas, Object Detection Using Yolo And Tensor Flow, Int. J. in Engi. Sci., 2024, Vol 1, Issue 1, 13-23. https://doi.org/10.5281/zenodo.11825059