International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 12, 2023
www.ijirse.com
ISSN: 2454-9665

Object Detection Using YOLO

Anubhav Sharma[1[21BCS8096]], Stayam Dhar Dwivedi [1[21BCS8024]]

Shafalii Sharma[1[E13752]]

[1] Chandigarh University, Mohali

lncs@springer.com

Abstract

Object detection in video streams depicts the core usage of computer vision, with its applications ranging from mask detection to surveillance systems. The You Only Look Once (YOLO) algorithm is famous for its high speed and accuracy. This review paper aims to provide review on recent advancements, challenges and applications of the YOLO algorithm. We also compare the performance for various YOLO algorithms like YOLO v7, v8 etc. Also, in this paper we explore the accuracy and efficiency of the YOLO algorithm by mapping to real world usage and testing protocols. It will also be tested in various environments but especially in video streams. The dynamicnature of YOLO algorithm will be tested.

Keywords: YOLO, Object Detection, Deep Learning, Computer Vision

1. Introduction

Real time object detection is turning out to be one of the best usage of deep learning for practical applications, it is applied in many fields including robotics, automated vehicles, AR, live video streams, dynamic weaponry and many more.[6] Of all the real time object detection algorithms, YOLO(you only look once)[7] stands out the most. It by far provides the highest accuracy and precision while also being open to all.

Throughout the history of YOLO it has seen many iterations, form v2 to v8 and getting better in each iteration. We will review the key innovations, ideas, performance changes, and also its functions to better understand this algorithm.

Not only that, but this paper aims to discuss and measure specific advancements of each version and see what we traded of in exchange for better speed and accuracy and also take into account the requirements of the system to handle such updates to the algorithm. Please note that this algorithm is still in development for furtherversions.[8]

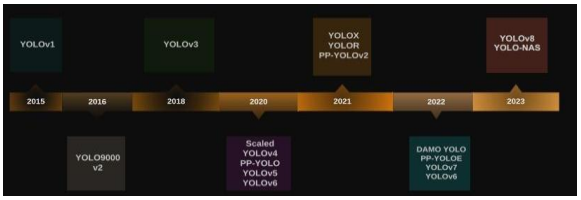


Fig 1: YOLO version Timelines.

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 12, 2023
www.ijirse.com
ISSN: 2454-9665

2. Evolution of YOLO

2.1. YOLOv1

This is what started it all. It worked by dividing an image into different part namely grid and then predicted the boundary boxes and class probabilities straight form an entire image. It worked well with good enough speed but had problems detectingsmall objects.

2.2. YOLOv2

This iteration of YOLO launched an year after v1, solved the most basic problem that was needed to be cleared. It solved the issue of detecting small objects by using anchor boxes. It also aimed to detect 9000 different object categories not present in training set. Hence it was also called YOLO9000.

2.3. YOLOv3

This is what started it all. It worked by dividing an image into different part namely grid and then predicted the boundary boxes and class probabilities straight form an entire image. It worked well with good enough speed but had problems detectingsmall objects.

2.4. YOLOv4

It further enhanced the speed and accuracy of object detection. It also used advanced techniques like CSPNet, PANet and SAM(Spatial Attention Module). Especially CSPNet was designed to enhance the learning capability of CNN(Convolutional neural networks).

2.5. YOLOv5

Released only a month after YOLOv4. It was developed by a company called Ultraytics, and claimed to have improved over its previous iterations. The original v5 model was not published as peer-reviewed research but instead as a Git Hub repository. The only improvement in this seemed to be the integration of anchor box selection process into the model.

2.6. YOLOv6

Released in June of 2022 it was considered the most improved version of the YOLO models at the time. It delivered very impressive results and provided excellent detection accuracy and speed. It is anchor free and uses Varifocal loss(VFL) and Distribution Focal loss(DFL).

2.7. YOLOv7

YOLOv6 while highly efficient and accurate, struggled with small objects due to using standard cross entropy loss function. YOLOv7 uses new loss function known as “focal loss”. It also has higher image resolution, almost double than YOLOv3 which allows for small objection detection.

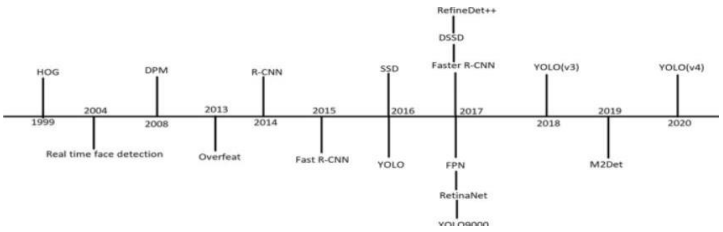


Fig 2.

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 12, 2023
www.ijirse.com
ISSN: 2454-9665

3. Architecture of YOLO

3.1. YOLOv1

YOLOv1 uses 24 convolutional layers followed up by 2 fully connected layers. Using grid on the image to divide it, i.e. dividing convolutional layers reduce the features space from preceding layers. The layers are pre-trained on Image-net classification tasks as 224x224 input image and then at 448x448 for detection.[1]

The system models detection as a regression problem. It divides the image into SxS grid and B bounding boxes and C for class probabilities. The formula makes up -

$S \times S \times (B * 5 + C)$[1] (1)

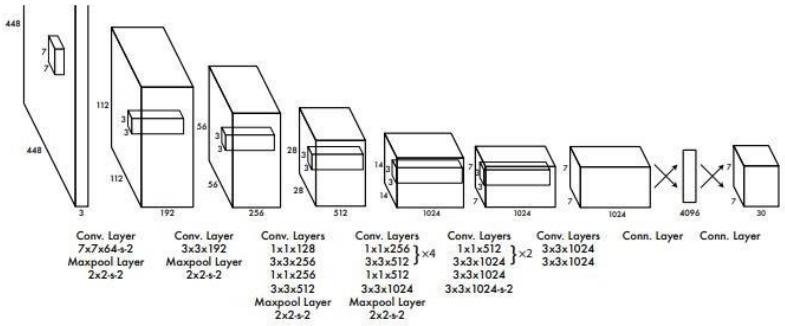


Fig 3. YOLOv1 architecture[1]

3.2. YOLOv2

YOLOv2 uses Darknet19(19 convolutional layers, 5 max pooling layers)[5]. In addition to this v2 uses anchor boxes, therefore instead of prediction arbitrary bounding box dimensions like YOLOv1 did, it predicts offsets to predefined anchor boxes.

It also uses batch normalization, high resolution classifier, convolutional with anchor boxes, dimensional clusters and direct location prediction.

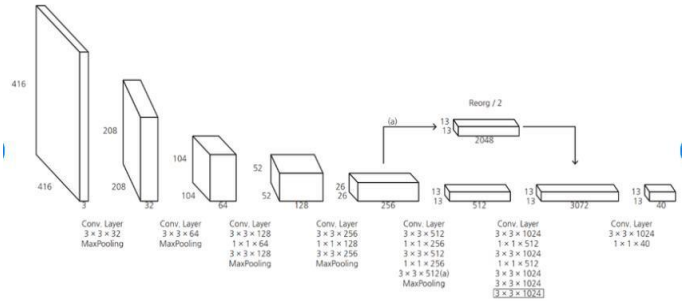


Fig 4. YOLOv2 Architecture[5]

3.3. YOLOv3

YOLOv2 uses Darknet53(52 convolutional layers, 5 max pooling layers). In addition to this v3 uses anchor boxes. It also uses Feature Pyramid Network(FPN). The anchor boxes are usually chosen by preforming

ISSN: 2454-9665

clustering algorithm of some kind of the training data set. This sysytem also predicts bounding boxes using dimension clusterswhere the prediction correspond to:[2]

$b_x = \sigma(t_x) + c_x$ (1)

$b_y = \sigma(t_y) + c_y$ (2)

$b_w = p_w e^{t_w}$ (3)

$b_h = p_h e^{t_h}$ (4)

Darknet-53 is depicted as follows:[2]

It also uses Bounding box prediction, class prediction, Predictions across scales,feature extractor.

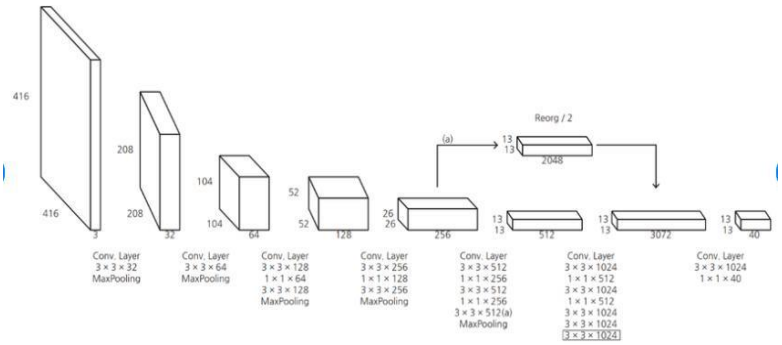


Fig 4. YOLOv3 Architecture[2]

3.4. YOLOv4

YOLOv4 breaks the norm of normal YOLO versions before it. See, the YOLO before it used to be one stage detected. Now form YOLOv4 Two stage detection is implemented to directly influence speed and accuracy. The parts of architecture are as follows: INPUT, BACKBONES,NECK [5] (additional blocks, path-aggregation blocks), HEADS(dense prediction, sparse prediction).

It also introduces bag of freebies and bag of specials. Bag of freebies are methods that only change the training strategy . Bag of specials are those plugin modules and post-processing methods that increase the cost by a small amount but increases the accuracy exponentially.

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 12, 2023
www.ijirse.com
ISSN: 2454-9665

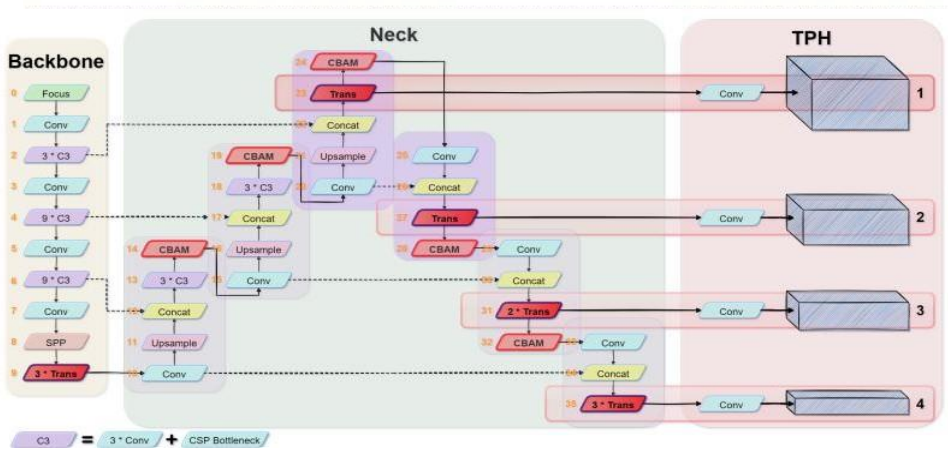


Fig 4. YOLOv4 Architecture[5]

3.5. YOLOv5

YOLOv5 generally uses the architecture of CSPDarknet53 with SPP layer as backbone,PANet as Neck and YOLO detection head[4]. Further optimization is provided by BOF and BOS.

	Type	Filters	Size	Output
	Convolutional	32	3 x 3	256 x 256
	Convolutional	64	3 x 3 / 2	128 x 128
1x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			128 x 128
	Convolutional	128	3 x 3 / 2	64 x 64
2x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			64 x 64
	Convolutional	256	3 x 3 / 2	32 x 32
8x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			32 x 32
	Convolutional	512	3 x 3 / 2	16 x 16
8x	Convolutional	256	1 x 1	
	Convolutional	512	3 x 3	
	Residual			16 x 16
	Convolutional	1024	3 x 3 / 2	8 x 8
4x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			8 x 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig 4. YOLOv5 Architecture[4]

Transformer Prediction Heads(TPH) are also integrated in YOLOv5. It helps to accurately localize objects in high density scenes like drone shots.[4] CBAM is also added to find region of interest. Also provides useful bag of tricks for streaming scenarios.

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 12, 2023
www.ijirse.com
ISSN: 2454-9665

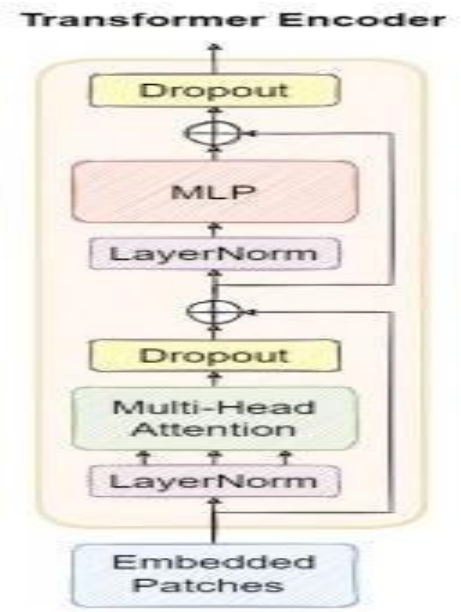
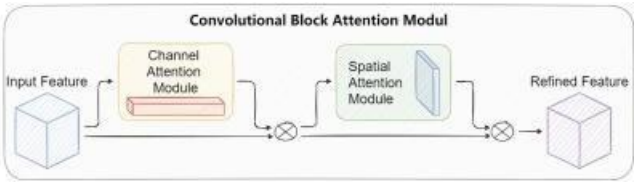


Fig 5.Transformer Encoder Architecture[4]

Fig 6. CBAM Architecture[4]

4. Conclusion

We saw the evolution of different YOLO algorithms and studied its architecture in great detail. The YOLO for



object detection is ever growing and the its possibilities are endless. With every iteration the YOLO grew and became better and better. Now there can be more modifications made to this system, which include but is not limited to :

1. Better Hardware
2. Better Optimization
3. Expansion to New Domains
4. Proliferation
5. Adaptability

And many more.

References

1. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhad: You Only Look Once: Unified, Real-Time Object Detection (2016). http://pjreddie.com/yolo/
2. Hoseph Redmon, AliFarhadi : YOLOv3: An Increamental Improvement(2018).

ISSN: 2454-9665

3. Alexey Bochkovskiy, Chein-Yao Wang, Hong-Yuan Mak Liao : YOLOv4: Optimal Speed and Accuracy of Object Detection(2020).
4. Xingkui Zhu1 * Shuchang Lyu1 * Xu Wang 1 Qi Zhao1:TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios.(2022)
5. Joseph Redmon, Ali Farhadi: YOLO 9000:Better, Faster, Stronger (2018). http://pjreddie.com/yolo9000/
6. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision, pages 740–755. Springer, 2014.
7. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database.In Computer Vision and Pattern Recognition, 2009. CVP IEEE Conference on, pages 248–255. IEEE, 2009.
8. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303– 338, 2010.
9. Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, Brian LeeA survey of modern deep learning based object detection models,