In object detection deep learning methods, YOLO shows supremum to Mask R-CNN

In object detection deep learning methods, YOLO shows supremum to Mask R-CNN

Shahriar Shakir Sumit[1*], Junzo Watada [1], Anurava Roy [1], DRA Rambli[1]

[1]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar 32610, Perak, Malaysia

Corresponding e-mail:- shahriar9121@gmail.com

Abstract. *Deep learning concept and algorithm play a pivotal role in solving various complicated problems such as playing games, forecasting economic future values, detecting objects in images. It could break through the bottle neck in conventional methods of neural networks and artificial intelligence. This paper will compare two influential deep learning algorithms in image processing and object detection, that is, Mask R-CNN and YOLO. Today, detection tasks become more complex when they come to numerous variations in the humans' perceived appearance, formation, attire, reasoning and the dynamic nature of their behaviour. It is also a challenging task to understand subtle details in their surroundings. For instance, radiance conditions, background clutter and partial or full occlusion. When a machine tries to interact with human or try to take pictures, it becomes hard for them to magnify the details of a human surrounding. In this study we have focused to detect humans effectively. The main objective of the present work is to compare the performance of YOLO and Mask R-CNN, which unveils the inability of Mask R-CNN in detecting tiny human figures among other prominent human images, and illustrate YOLO was successful in detecting most of the human figures in an image with higher accuracy. Therefore, the paper evaluates and differentiates the performance of YOLO from the deep learning method Mask R-CNN in two points, (1) detection ability and (2) computation time. Since, the machine learning algorithms are mostly data specific, the authors believe that the presented results might vary with the varying nature of the data under observation. In another way, the presented data might be seen as a counter example of unveiling the detection inaccuracy of the Mask R-CNN.*

Keywords—*Object detection, CNN, Mask R-CNN, YOLO*

1. Introduction

It is easy for humans to understand pictures, where they can recognize objects in the pictures by twinkling of an eye without any trouble. They can count the number of people in the images and can guess even the human's emotional condition such as happiness, sadness and anger. Moreover, they can predict their relationship whether they are family members, friends group, colleagues or just pedestrians in a walk.

1.1 Research Background

The computer vision is expected to provide computer-support for this human ability [20]. The aim of Computer Vision is to let a computer understand an image or video by the eyes of a camera. The

computer vision is an area that involves various techniques for acquiring, processing, analyzing, and understanding images. To make decisions it creates numerical or symbolic information from high-dimensional data in the real world [20].

Object detection is considered one of the most core responsibilities in computer vision. Object detection is the method of identifying occurrences in pictures and videos of a particular type of objects (e.g. pedestrians, cars and bikes). Human detection has been extensively used in various applications, especially in intelligent video surveillance systems [3], person recognition [4], person re-identification [5], body part tracking [6], action recognition [7], etc. as a hot topic in the field of computer vision, for example, Kalayeh *et al*. [5] developed SPReID for human re-identification. To make computer vision algorithms convenient in the real world for detection they must be executed in a real-time process.

A computer vision method should apply easy, rational but stable as well as quick computer vision techniques to fulfill the requirements of the real world. Recently several methods have achieved better performance to detect humans object class [21] but not in high standard to be satisfied. As the image resolution increases, many advanced computer vision algorithms are unable to accomplish real-time efficiency and are unable to manage the more complicated assignments [9] for instance crowed scenes. It is therefore extremely desirable to have a real-time system capability of handling accurately and effectively high-resolution pictures and complicated environments.

1.2 Recent Literature

Object detection is one of the key computer vision tasks. Detecting as well as locating an object class in the images or videos is the aim of object detection. In computer vision area detecting objects especially humans class is one of the most significant challenges. A lot of methods have been proposed till now for object detection, for example, HOG (Histogram of oriented gradients) [16]. The actual object detection is started from 2001 when Paul Viola and Michael Jones proposed a feature based framework with high accuracy and speed is considered as the first object detection framework [23]. This approach is highly efficient but cannot handle the complexity of images and videos in real world [22]. One of the most popular object detection methods Histogram of Oriental Gradients (HOG) descriptor has been proposed in 2005 by Dalal and Triggs [16]. Histogram of Oriented Gradient feature vectors are extracted where the detector window is tiled with a grid of overlapping blocks [16]. For classification of objects/non-objects the combined vectors are provided to a linear SVM. A histogram of the gradient orientations and magnitudes represents each image cell hence HOG is more stable to light, shadows, etc. [16]. After that many upgraded methods have been offered built on HOG features that have been broadly used to detect pedestrian class. For an example the combination of Local Binary Pattern (LBP) feature and HOG features framework that has been proposed by Wang and Han [26]. In combination with HOG color self-similarity termed CSS which has been introduced by Walk *et al*. increases state-of-the-art classification performance for not only static images but also image sequences [24]. Another fast detection technique has been proposed nearly the same performance as the HOG detector in real time by Dollar *et al*. [25]. There are some other popular object detectors: SIFT (Scale-invariant feature transform), this technique identifies and explains the local characteristics in pictures [18], SURF (Speeded up robust features), it is a static Hessian scale and rotation based descriptor which extracts interested attributes in the images [17], BOG (Bag-of-words), this model deals with the picture characteristics as words [19] and their variants. An object detection system built on mixtures of multiscale deformable part models (DPM) has been described in [15]. This system can illustrate highly variable classes of objects and gain state-of-the-art results in the challenges of PASCAL object detection. This model is divided by two classes. In first class it involves a star-structured part-based model which is characterized by a “root” filter plus a set of part filters and deformation models. By a mixture of star models, it signifies an object category in its second class. Using latent SVM these parts are trained on multi-scale HOG features [15].

In recent years the field of computer vision has been revolutionized by deep learning techniques after the publication of the AlexNet Convolutional Neural Network Model [10]. Deep learning methods are currently state-of-the-art in computer vision and image processing problems, particularly object detection [11] due to the availability of labeled image data sets with millions of images [14] and

computer hardware that allow robustness in computing. Deep Convolution Neural Networks (CNNs) have demonstrated remarkable performance in different vision responsibilities appreciations to deep networks (e.g. ZF Net, VGG, Inception/GoogLeNet, ResNet) and Regions with CNN features (RCNN) [12]. Besides we can separate deep learning methods by regression/classification-built detectors also called as single stage detectors and region proposal-made detectors also named as multi-stage detectors. Multi-stage detectors consist of several connected stages, taking in the generation of regional proposals, CNN extraction features, classification, and bbox (bounding box) regression, which are generally trained individually. Region proposal-based detectors include mostly SPP-net (spatial pyramid pooling) [27], FPN (Feature pyramid networks) [28], R-FCN (region-based fully convolutional network) [29], R-CNN [12], Fast R-CNN [30], Faster R-CNN [31], Mask R-CNN [1]. Single stage detectors constructed with inclusive regression/classification can minimize computational time by mapping directly from picture pixels to bbox (bounding box) coordinates and frequency probabilities. Regression/classification-built detectors mostly include YOLO [2], SSD (single shot multi-box detector) [32], DSSD (deconvolutional single shot detector) [12], DSOD (deeply supervised object detectors) [12], G-CNN [12], AttentionNet [12], Multibox [12].

Table 1. An overview of single stage and multistage object detectors

Method name	Learning paradigm	Proposal	Softmax function	Framework	Programming language
R-CNN[12]	SGD, BP	Selective search	Used	Caffe	MATLAB
Fast R-CNN[30]	SGD	Selective search	Used	Caffe	Python
Faster R-CNN[31]	SGD	RPN	Used	Caffe	Python/ MATLAB
Mask R-CNN[1]	SGD	RPN	Used	TensorFlow/Keras	Python
SPP-net[27]	SGD	Edge Boxes	Used	Caffe	MATLAB
R-FCN[29]	SGD	RPN	not used	Caffe	MATLAB
YOLO[2]	SGD	-	not used	Darknet	C
SSD[32]	SGD	-	Used	Caffe	C++
FPN[28]	Synchronized SGD	RPN	used	TensorFlow	Python

The recently introduced ImageNet [13] task for video object detection (VID) brings the object detection task into the video domain, where the locations of objects in each frame must be annotated.

The remaining part of this paper is described as follows: Part 2 describes the Used Methodologies in Brief, Part 3 explains the Result and Discussion and Part 4 concludes the paper.

2. Used Methodologies in Brief

This section briefly touches upon the basic concepts of Mask R-CNN and YOLO as follows:

2.1 *Mask R-CNN* [1]

Mask R-CNN (regional convolutional neural network) is the extended version of Faster R-CNN. Mask R-CNN has same two stage: the first phase scrutinizes the picture and produces proposals. In the second phase it categorizes the proposals and results in bounding boxes and masks. Feature Pyramid Network (FPN) and a ResNet101 are used as backbone architecture in Mask R-CNN. To anticipate segmentation masks on each Region of Interest (RoI) a branch is added which is used for classification and bounding box regression simultaneously with the existing branch. ROI pooling denotes cropping and re-sizing a part of a feature map to a rigid size. Moreover, pixel-to-pixel alignment is the key elements of Mask R-CNN which is absent in Fast/Faster R-CNN. For particular instance of an object in the picture, the model produces bounding boxes and segmentation masks.

2.2 *YOLO* [2]
"You Only Look Once" (YOLO) is one stage object detector that can predict a definite object on each region of the feature maps deprived of the cascaded location classification phase. YOLO employs a particular CNN network and applies bounding boxes to classify and locate the object. It splits the picture into a grid of S×S; $S \in \mathbb{Z}^+$. It would be the object if an object's focal point swoops a grid cell. YOLO Being a single stage detection system, YOLO is immensely fast and is being used in real time for object detection.

3. Result and Discussion

Before going to further details of the experimentation, it would be apt to show further details of the used data set and system specification as follows:

3.1 Data-Set Specification

Fig. 1. Several sample images from our custom data-set.

Without using data-sets for example COCO, PASCAL VOC we have tested the efficiency of these two models and found YOLO is superior in our data-set. A part of the used data set is shown in Fig. 1. The data-set characteristics are as follows:

1. Individual images contain random human figure(s) both overlapping and non-overlapping;
2. Each image is of size 400 X 600 pixels
3. Data set contains 500 images (including images from internet containing human figures, Asian, Indian, European, African) in different geographic locations with different backgrounds.

3.2 System Specification
All experimental verifications are done using GTX 1060 GPU with RAM 64GB, multi-threaded CPU with 8 logical CPUs with L1 cash 256KB, L2 cash 1MB, L3 cash 6MB with base OS image of Ubuntu-16.04 LTS and TensorFlow framework 1.11.0, Keras 2.2.4 and Python 3.6.3.

3.3 Experimental Result
Mask R-CNN is a multi-stage strategy that fails to detect all human figures present in a specific image at the beginning (see Fig. 2). On contrary, YOLO satisfactorily removes the drawback, evidences are shown in Fig. 3.

Note that: the human figure in the red box is identified at the first stage, while in the next stage the blue box was identified.

Fig. 2. Multistage human detection using Mask R-CNN.

Note that: YOLO simultaneously detects all the images. It can detect all the human figures in an image.

Fig. 3. Human detection using YOLO.

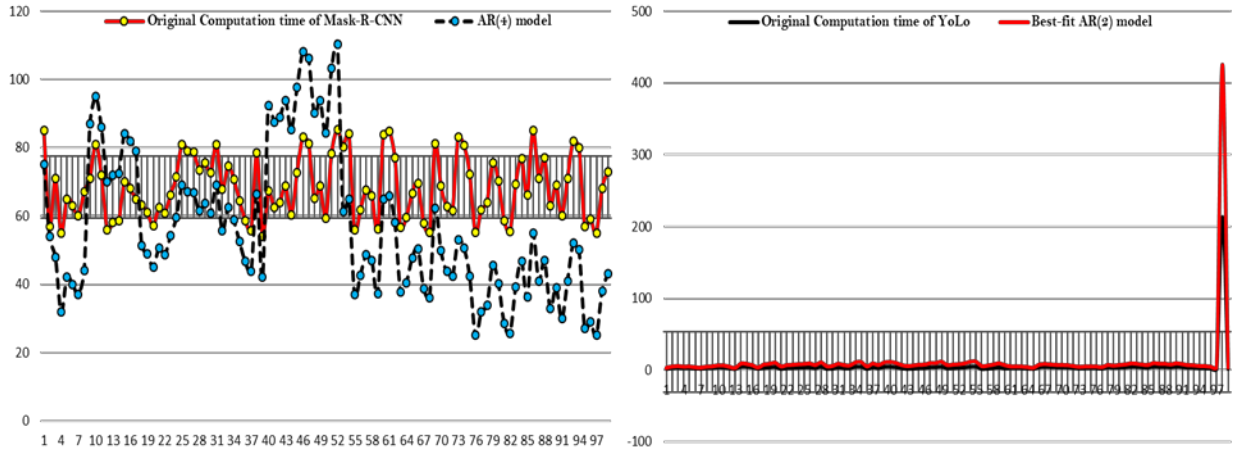
Note that: Mask R-CNN shows 67.63215ms as average, whereas YOLO maintains much better average computational time.

Table 2. Computational time for Mask R-CNN and YOLO

Sample no.	Mask R-CNN (ms)	Mean	YOLO (ms)	Mean
n = 1	85		2	
n = 2	57		3	
n = 3	71		3.3	
n = 4	55		2.9	
n = 5	65	67.63215	3	5.48544
n = 6	63		2.5	
...	...		...	
...	...		...	
...	...		...	
n=100	73		2	

Fig. 2 shows Mask R-CNN consists of a multi-stage strategy. At the first stage Mask R-CNN can detect one human figure from the image. In the next stage, Mask R-CNN detects another human figure, whereas YOLO detects all the human figures at the very first attempt shown in Fig. 3. Quite surprisingly, Mask R-CNN is found unable to detect few human figures in a specific image, whereas no such problem occurs in case of YOLO. Undetected human figure (using Mask R-CNN) in Fig. 2 is indicated with a blue circle in, whereas no such anomaly object was found in case of YOLO (evidences shown in Fig. 3). In our experiments, we found YOLO's computational time is shorter than Mask R-CNN. Details of

computational time are shown in Table 2. We further continue our investigation to compare the execution time of Mask R-CNN and YOLO over 100 randomly selected batches (both overlapping and non-overlapping), each of which contains 20 random images, taken from our data repository.



Note that: (1) Shaded regions in both the figures show the standard deviation of the original computational times of Mask R-CNN and YOLO. (2) Left graph shows the computational time of Mask R-CNN execution for randomly selected 100 samples from our data repository along with the best-fit AR (5) model (selected using AIC and BIC values). The right one is the visualization of the YOLO execution over the same 100 random samples along with the best-fit AR (2) model.

Fig. 4. Computation time comparison.

We have estimated the curves using univariate models and found the best-fit curves named AR (5) and AR (2) respectively for Mask R-CNN and YOLO models. It is evident from the result that the order of best-fit AR for Mask R-CNN is higher than that of YOLO and involves more historical data than that of YOLO. Hence, it can be concluded that the time complexity of Mask R-CNN is higher than that of YOLO.

4. Conclusion

Both Mask R-CNN and YOLO can detect object. Mask R-CNN can benefit from extra data, even if that data is unlabeled. Mask R-CNN is also capable for instance segmentation. It can be used in Human pose approximation. The key findings of this study can be concise as comparing to YOLO, Mask R-CNN takes more time for detection. Moreover, in few cases Mask R-CNN was found unable to detect human figures, one instance has been presented in Fig.2 (Section 3), whereas YOLO is capable of detecting all types of human figures. YOLO can be used in any kind of object detection in real time and can be considered as the better model between the mentioned two. Obtained results are data specific and might change with changes in data distribution.

References

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R- CNN,” In International Conference on Computer Vision, 2017.
[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” In Conference on Computer Vision and Pattern Recognition, 2016.
[3] Jong Sun Kim, Dong Hae Yeom, Young Hoon Joo, “Fast and robust algorithm of tracking multiple moving objects for intelligent video surveillance systems,” In IEEE Trans on Consumer Electronics, vol. 57, no. kk, pp. mmm-nnn2011.

[4] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," In IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 8, pp. 1114–1127, 2008.
[5] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, M. Shah, "Human Semantic Parsing for Person Re-identification", In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062–1071, 2018.
[6] Jain Stoble B., Sreeraj M., "Multi-posture Human Detection Based on Hybrid HOG-BO Feature," In Fifth International Conference on Advances in Computing and Communications (ICACC), pp. 37-40, 2015.
[7] R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic Petri net framework for human activity detection in video," In IEEE Trans. Multimedia, vol. 10, no. 8, pp. 1429–1443, 2008.
[8] D.M.P.Felzenszwalb, R.Girshick, "Cascade object detection with deformable part models," In Conference on Computer Vision and Pattern Recognition, 2010.
[9] K. S. Omar Javed and M. Shah, "Automated surveillance in realistic scenarios," In IEEE MultiMedia, vol.14, pp. 30-39, 2007.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In 26th Annual Conference on Neural Information Processing Systems, pp. 1106–1114, 2012.
[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," In Conference on Computer Vision and Pattern Recognition, 2014.
[12] ZQ Zhao, P Zheng, S Xu, and X Wu, "Object Detection with Deep Learning: A Review," In IEEE Trans. on Neural Networks and Learning System, 2019.
[13] K. Kang, W. Ouyang, H. Li, X. Wang, "Object Detection from Video Tubelets with Convolutional Neural Networks," In Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," In International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.
[15] Felzenszwalb,P.F., Girshick,R.B., McAllester,D., "Ramanan,D..Object detection with discriminatively trained part-based models," In IEEE Trans. on PAMI., vol. 32, iss. 9, pp. 1627–1645, 2010.
[16] Dalal, N., Triggs, B., "Histograms of oriented gradients for human detection," In International Conference on computer vision & Pattern Recognition, vol. 1, pp. 886–893, 2005.
[17] Bay, H., Tuytelaars, T., Van Gool, L., "Surf: Speeded up robust features," In European conference on computer vision, pp. 404–417, 2006.
[18] Lowe, D.G., "Object recognition from local scale-invariant features," In International Conference on Computer Vision, vol. 99, pp. 1150–1157, 1999.
[19] Fei-Fei, L., Perona, P., "A bayesian hierarchical model for learning natural scene categories," In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 524–531, 2005.
[20] Reinhard Klette, "Concise Computer Vision". Springer, 2014.
[21] X. Wang, G. Hua, and T. X. Han. "Detection by detections: Non-parametric detector adaptation for a video". In Conference on Computer Vision and Pattern Recognition, 2012.
[22] Liao, S., Jain, A.K., Li, S.Z., "A fast and accurate unconstrained face detector," In IEEE Trans. on pattern analysis and machine intelligence, vol. 38, no. kk, pp. 211–223, 2015.
[23] P. Viola; M. Jones. "Rapid Object Detection Using a Boosted Cascade of Simple Features," In Conference on Computer Vision and Pattern Recognition, pp. 511–518, 2001.
[24] S. Walk, N. Majer, K. Schindler, and B. Schiele. "New features and insights for pedestrian detection". In IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[25] P. Dollar, S. Belongie, and P. Perona. "The fastest pedestrian detector in the west". In British Machine Vision Conference, 2010.

[26] X. Wang and T. Han. "An hog-lbp human detector with partial occlusion handling". In International Conference on Computer Vision, 2009.
[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," In IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.
[28] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, "Feature pyramid networks for object detection," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944, 2017.
[29] Jifeng Dai, Yi Li, Kaiming He, Jian Sun, "R-FCN: Object detection via region-based fully convolutional networks," In Conference on Neural Information Processing Systems (NIPS), Advances in Neural Information Processing Systems, pp. 379–387, 2016.
[30] R. Girshick, "Fast R-CNN," In IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, 2015.
[31] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," In Conference on Neural Information Processing Systems (NIPS), Advances in Neural Information Processing Systems, pp. 91–99, 2015.
[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single shot multibox detector," In European Conference on Computer Vision (ECCV), pp. 21–37, 2016.