Review on NMS-Free Object Detection Frameworks using YOLO

Md Firoz Alam(2023PIS5097)
Submitted to
Prof. Girdhari Singh

Abstract

Real-time object detection has always been a focal point of research in the area of computer vision, which aims to accurately predict the categories and positions of objects in an image under low latency. Traditional object detection approaches often rely on Non-Maximum Suppression (NMS) to eliminate redundant bounding boxes in overlapping detections. However, NMS has inherent limitations, particularly in scenarios with densely packed objects. This report examines recent advancements in object detection frameworks that eliminate the need for NMS, focusing specifically on models inspired by YOLO (You Only Look Once). It discusses the challenges associated with detecting objects in dense environments and highlights alternative methodologies aimed at enhancing detection accuracy and computational performance without NMS. The objective of this review is to throw light on the mechanisms and benefits of these NMS-free frameworks.

Index Terms

Object detection, NMS-free, deep learning, YOLO, accuracy, efficiency, dense object environments.

I. Introduction

In many computer vision applications, including robots, autonomous driving, and video surveillance, object detection is essential. Conventional object identification techniques rely on algorithms such as SSD, Faster R-CNN, and YOLO. These techniques generate numerous people on foot and then utilize Non-Maximum Suppression (NMS) to filter out unnecessary bounding boxes. NMS performs poorly in dense object settings, when objects are close to one another, even while it performs well in sparse environments.

NMS frequently misses detections or has several overlapping bounding boxes in these circumstances because it is unable to discriminate between objects that are closely packed together. This problem is especially troublesome in situations like congested metropolitan settings, where a lot of things, including bicycles, cars, and pedestrians, might be grouped together and pose a serious risk.

To overcome these obstacles and raise detection accuracy in crowded environments, NMS techniques were learned, along with extensions and substitutes like soft-NMS and adaptive-NMS.By altering the way overlapping boxes are suppressed or by adding contextual information about the relationships between the objects, these techniques seek to maintain more accurate detections. Nevertheless, object recognition in intricate, crowded surroundings is still a challenging task that calls for constant research and development.

II. Background and Related Work

Several advancements have been made in the field of object detection, specifically with regard to eliminating the reliance on NMS. In traditional YOLO models, NMS is used to select the most accurate bounding box from multiple overlapping predictions. However, this process can suppress correct detections and increase computational complexity in dense scenarios below is given related work which has already been done.

1) **Real-time object detectors.** Real-time object detection aims to classify and locate objects under low latency, which is crucial for real-world applications. Over the past years, substantial efforts have been directed towards developing efficient detectors. Particularly, the YOLO series [8, 6, 7, 1, 3, 5, 10, 4, 11] stand out as the mainstream ones.
2) **End-to-end object detector:** It is a CV technique that uses a single stage to perform both object detection and classification as a paradigm shift from traditional pipelines, offering streamlined architectures [9]. DETR [2] introduces the transformer architecture and adopts Hungarian loss to achieve one-to-one matching prediction.

   Furthermore, studies like those by Liu et al. [11] and Wang et al. [1] have proposed variations of YOLO that incorporate techniques like attention mechanisms or alternative loss functions to reduce overlap between predictions, improving both precision and recall.

Author	Title	Aim	Methodology	Performance	Limitations
Redmon et al.2016[8]	You Only Look Once: United, Real-Time Object Detection.	to show that YOLO is a quick, precise, and broadly applicable object detection technique that works well with a variety of datasets and real-time applications.	YOLO presents a unified CNN-based object detection system that predicts bounding boxes and class probabilities in a single pass, treating detection as a single regression problem. High-accuracy real-time processing and streamlined end-to-end training are made possible by this approach.	While maintaining competitive accuracy and twice the mAP of existing real-time detectors, YOLO processes images in real-time at 45 frames per second, while Fast YOLO achieves 155 frames per second.	Limited number of objects detected, small objects not properly detected, struggle to localize small objects with fine details, loss function design, incorrect localization of bounding box.
Redmon et al. 2022[6]	YOLO9000: Better, Faster, Stronger	intends to present YOLO9000, a cutting-edge real-time object identification technology that can identify more than 9000 object categories.	In order to improve detection accuracy, the study presents YOLO9000 with improvements like anchor boxes, multi-scale training, and fine-grained features. In order to identify more than 9000 object types, it also employs joint training using classification and detection datasets using a hierarchical structure (WordTree).	AP: 76.8 at 67 FPS (real-time), 78.6 at 40 FPS (high accuracy),mAP: 73.4, competitive with Faster R-CNN and SSD	Geographical limitations, imbalanced loss functions, limited detection data for some classes, trade-offs between speed and accuracy, and localization errors.
Redmon et al. 2018[7]	YOLOv3: An Incremental Improvement	to create a bigger network (YOLOv3) that performs quickly and is more accurate than earlier iterations.	It uses a powerful feature extractor (Darknet-53) in conjunction with a multi-scale prediction technique to improve detection speed and accuracy. For bounding box and class predictions, it makes use of logistic classifiers, anchor boxes, and binary cross-entropy loss.	RetinaNet achieves 57.5 AP50 in 198 ms, showing similar performance but is 3.8× slower than YOLOv3	Reduced Resolution for Larger Models, Limited Evaluation Metrics, and Performance on Smaller Objects.

TABLE I: Summary of Related Work

Author	Title	Aim	Methodology	Performance	Limitations
Alexey et al.2016[1]	YOLOv4: Optimal Speed and Accuracy of Object Detection	the creation of a quick and precise object identification model (YOLOv4) that is tailored for real-time use on traditional GPUs.	CSPDarknet53 for efficient feature extraction,Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN) for feature aggregation.	With a real-time inference performance of about 65 frames per second on a Tesla V100 GPU, YOLOv4 achieves an average precision of 43.5% and an AP50 of 65.7% on the MS COCO dataset. At 10% accuracy and 12% speed, it surpasses YOLOv3, and it is on par with cutting-edge detectors such as EfficientDet.	Even while YOLOv4 is fast and accurate, it might not be able to compete with state-of-the-art anchor-free models for tasks that call for incredibly accurate localization or segmentation.
Aduen et al. 2022[6]	YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles.	For applications involving autonomous racing, the YOLOv5 object detector's small-object detection is improved. Improving small object detection performance without appreciably lengthening inference time is the aim.	DenseNet was used in place of YOLOv5's default backbone in order to enhance feature retention and enhance small object detection.	up to 6.9% improvement in mAP for small objects at 50% IoU,An average 2.7% overall mAP improvement across all scales	The study's generalizability to other applications is limited by its concentration on a cone dataset. Enhancements in performance cause a modest increase in inference time, which affects systems with limited resources. Validation requires more extensive testing on a variety of datasets.
Chuyi Li et al. 2018[5]	YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications.	The goal is to create YOLOv6, a new object identification framework that is ideal for industrial applications and excels in speed and accuracy.	incorporated an efficient decoupled head and a modified PAN topology for the neck, and created separate backbones for the small and big models (EfficientRep and CSPStackRep Block, respectively).	The new benchmark for accuracy-speed trade-offs was set by Quantized YOLOv6-S, which achieved 43.3 AP at 869 FPS.	Deployment: The flexibility to use different hardware is limited by YOLOv6's,Generalization: On non-GPU machines, quantization techniques might not work as well,Complexity: Larger models are more difficult to use in circumstances with limited computational resources.

III. Problem Statement

The problem addressed in this work is the inefficiency and potential inaccuracy introduced by NMS in object detection, particularly in dense environments. Objects in such environments often overlap, and NMS can eliminate valid detections by discarding low-confidence bounding boxes that might represent different objects. Additionally, the computational cost of applying NMS increases as the number of detections grows, making it less suitable for real-time applications.

Considering that we concentrate on methodologies based on the YOLO architecture, our objective is to explore and analyze NMS-free object recognition frameworks that do not require post-processing procedures like NMS. The aim is to enhance recognizing items efficiency and accuracy in occupied and heavily populated regions.

IV. Methodology

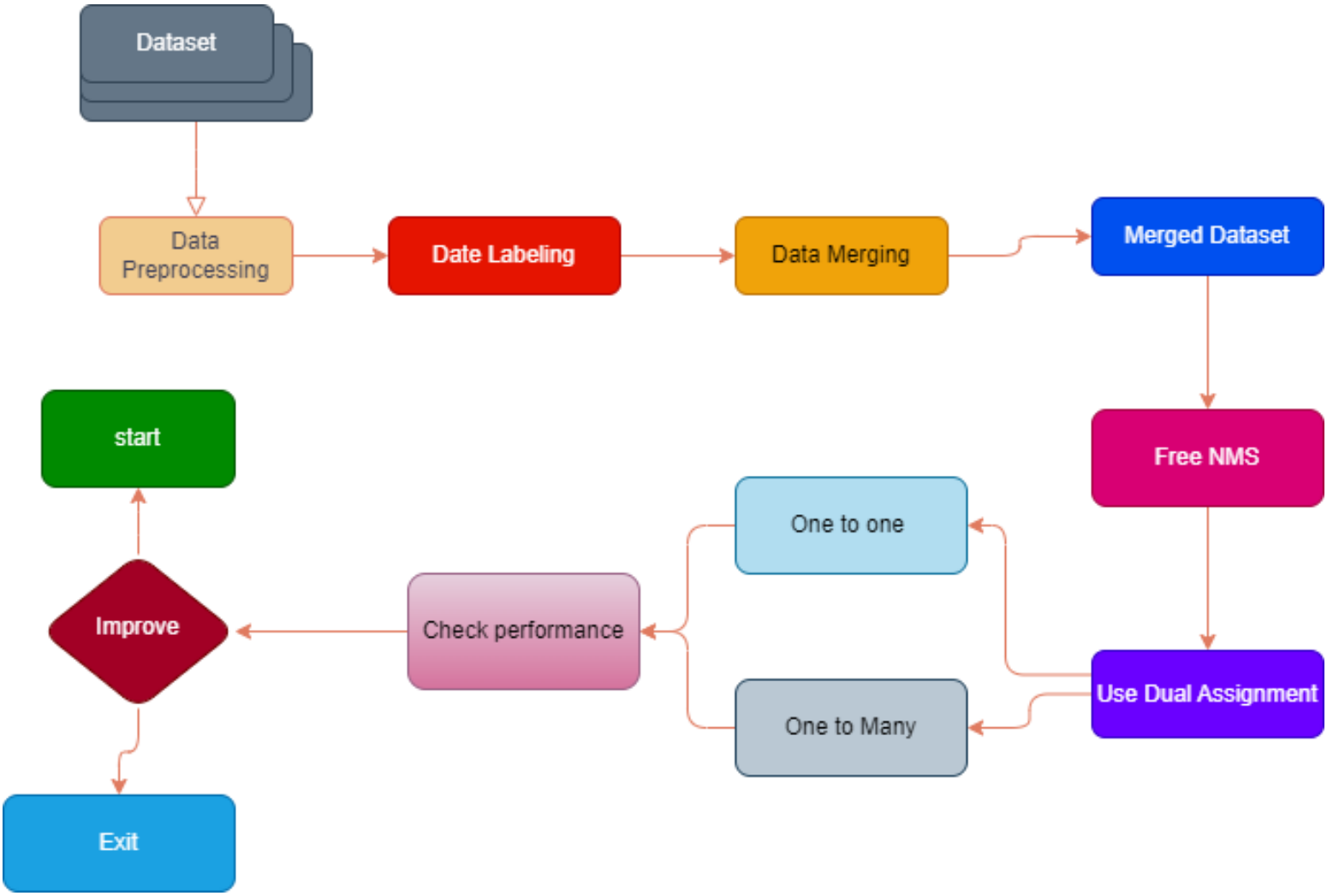

Fig. 1: Flowchart describing the proposed methodology.

A. Framework Overview

In an effort to directly predicted bounding boxes without overlapping boxes, the NMS-free YOLO-based frameworks discussed in this review modify the traditional YOLO architecture. Post-processing procedures, loss functions, and multiple network designs are utilized to achieve this. Employing anchor-free models is a particular approach for which the network predicts the centers of items or the corners of boxes with boundaries.
By ensuring that each component discovered corresponds to a distinct bounding box, these models can be configured suitable for handling dense environments without the need for post-processing suppression. Another approach exploits methods of attention that assign precedence to more accurate predictions by emphasizing on regions of the image with a high product density.
This reduces the necessity for NMS by enabling the model to differentiate between products which are tightly grouped closely. Another technique utilizes distance-based loss functions, which encourage the network to prepare for well-separated bounding boxes and decrease overlap between objects that are understood. For improved localization by avoiding redundant boxes in crowded circumstances, graph-based strategies were additionally developed. These algorithms are used to specify the connections between discovered objects..

B. Model Design

The following layers constitute an NMS-free YOLO-based model's normally architecture:

- Backbone CNN: The input image's features are obtained utilising a CNN which already has been trained, such as ResNet or EfficientNet. The previously FPN, or Feature Pyramid Network: Multi-scale feature extraction may be addressed by an FPN, which makes it especially beneficial with observing small objects in crowded environments.
- Objectness Prediction:The likelihood for an object is going to be present in a particular region of the image is determined with the network.
- Bounding Box Regression: Using a loss function that penalizes overlap, the network learns to predict non-overlapping bounding boxes directly rather than through NMS.This architecture allows the model to predict bounding boxes in a manner that reduces overlap between detections without the need for post-processing.

C. Training and Evaluation

The model uses a combination of cross-entropy loss for classification and a customized overlap penalty loss for improving bounding box predictions when it is trained on datasets with densely packed objects, like COCO or ADE20K. Standard benchmarks such as Mean Average Precision (mAP) and Intersection over Union (IoU) are used in the assessment. The evaluation results are contrasted with those of traditional YOLO frameworks that use NMS. Performance metrics like recall, precision, and processing time during inference are used to assess the effectiveness of NMS-free detection systems.

V. Comparision of different yolo model

VI. Conclusion

Several Yolo-based object detection frameworks were analyzed in this paper, with an emphasis on those based on the YOLO architecture [8, 6, 7, 1, 3, 5]. We investigated how these frameworks improve detection accuracy and

sno	Detector	No of n-layers	FLOPS	FPS	mAP	Used Dataset
1	Yolov1	26	8.65	45	63.5	Voc Dataset
2	yolov1-Tiny	9	6.65	155	52.8	Voc Dataset
3	yolov2	32	62.95	40	48.2	Coco Dataset
4	yolov2-Tiny	16	5.42	244	23.6	Coco Dataset
5	yolov3	106	140.7	20	57.8	Coco Dataset
6	yolov3-Tiny	24	5.57	220	33.2	Coco Dataset

TABLE II: Comparison of YOLO models

computing efficiency by overcoming the drawbacks of Non-Maximum Suppression NMS-free in cases with dense objects. According to the analysis, NMS-free models often perform better than conventional NMS-based techniques, exhibiting notable gains in processing speed, precision, and recall. In addition, these models perform better in complex situations. where things are crammed together.

In order to enhance detection performance in increasingly difficult and congested settings and broaden their applicability to a greater variety of real-world tasks and applications, future research will seek to further optimize these methods. Furthermore, attempts will be made to incorporate these models into real-time systems, confirming their durability and scalability in dynamic settings.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection". In: *arXiv preprint arXiv:2004.10934* (2020).

[2] Nicolas Carion et al. "End-to-End Object Detection with Transformers". In: *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[3] Glenn Jocher. *YOLOv5 Release v7.0*. https://github.com/ultralytics/yolov5/tree/v7.0. 2022.

[4] Glenn Jocher. *YOLOv8*. https://github.com/ultralytics/ultralytics/tree/main. 2023.

[5] Chuyi Li et al. "YOLOv6 v3.0: A Full-Scale Reloading". In: *arXiv preprint arXiv:2301.05586* (2023).

[6] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.

[7] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[8] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[9] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. "End-to-End People Detection in Crowded Scenes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2325–2333.

[10] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7464–7475.

[11] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information”. In: *arXiv preprint arXiv:2402.13616* (2024).