SHODH SAGAR®
International Journal for Research Publication and Seminar
ISSN: 2278-6848 | Vol. 15 | Issue 3 | Jul - Sep 2024 | Peer Reviewed & Refereed

Object Detection in Unstructured Driving Environments

Shrey
2021562
shrey21562@iiitd.ac.in

Vasu Kapoor
2021573
vasu2003kapoor@gmail.com

Vinayak Sharma
2021574
vinayak21574@iiitd.ac.in

DOI: https://doi.org/10.36676/jrps.v15.i3.1459

Accepted:14-08-2024 Published: 17-08-2024

***C**orresponding auth

Abstract

This paper conducts a comprehensive error analysis of the inference process performed on the YOLOv8 and RTDETR model, utilizing two distinct datasets: MS COCO, on which YOLOv8 and RT-DETR is originally trained, andIDD, a separate dataset. The primary focus lies on evaluating model performance using mean Average Precision (mAP) and Intersection over Union (IoU) metrics. Through rigorous experimentation and analysis, we investigate the discrepancies in model performance when applied to thesediverse datasets. The findings shed light on the strengths and weaknesses of the YOLOv8 and RT-DETR model across different data domains, offering valuable insights for improving object detection systems in real-world applications.

1. Problem Statement

Despite the widespread adoption of object detection modelslike YOLOv8 and RT-DETR, there remains a critical needto understand their performance variations across differentdatasets. This paper aims to address this gap by conductingan error analysis of the YOLOv8 and RT-DETR model's inference on two distinct datasets: MS COCO, the dataset onwhich YOLOv8 and RT-DETR is trained, and IDD, a different dataset. The specific focus is on evaluating model performance using mean Average Precision (mAP) and Intersection over Union (IoU) metrics. By identifying and analyzing the discrepancies in model performance across these datasets, this study seeks to provide insights into the model's effectiveness and limitations in real-world scenarios.

2. Literature Review

2.1. YOLO

YOLOv8[5], utilizes a deep neural network with numerous convolutional layers, including backbone networks like CSPDarknet53 and SPP (Spatial Pyramid Pooling), followed by detection layers. YOLOv8 aims to strike a balance between speed and accuracy, crucial for real-time applications. It achieves this by optimizing various components of the network, including backbone architecture,

training strategies, and post-processing techniques. YOLOv8 introduces optimizations to enhance inference speed without compromising accuracy. Techniques such as model pruning, network quantization, and efficient post-processing are employed to achieve real-time performance on resourceconstrained devices. The paper provides comprehensive experimental results on benchmark datasets, demonstratingthe superior performance of YOLOv8 compared to previousversions and other state-of-the-art object detection models in terms of both speed and accuracy.

2.2. RT-DETR

RT-DETR[2], a groundbreaking object detector developed by Baidu, combines Vision Transformers (ViT) with innovative techniques to achieve real-time performance without compromising accuracy. Its efficient architecture processes multiscale features by separating intra-scale interaction and cross-scale fusion, reducing computational costs and enabling rapid detection. Notably, it features IoU-aware query selection for improved object detection accuracy and supports flexible adjustment of inference speed through decoder layer modifications, making it highly adaptable for diverse real-time scenarios. Compatible with accelerated backends like CUDA with TensorRT, RT-DETR surpassesmany existing real-time detectors in performance. It beats YOLO in terms of performance.

3. Dataset Description

3.1. MS COCO 2017

The Microsoft Common Objects in Context (COCO) 2017dataset is a pivotal resource in computer vision, comprising over 330,000 meticulously annotated images covering 80 object categories with segmentation masks and bounding boxes. Renowned for its diversity and high-quality annotations, it offers a robust testbed for object recognition
tasks. With its broad spectrum of object types and scenes, COCO challenges models to generalize effectively. Its support for multiple tasks including object detection, instance segmentation, and image captioning fosters comprehensive research and development. Through annual challenges and a permissive license, COCO encourages innovation, collaboration, and reproducibility, making it an invaluable asset inadvancing the frontiers of computer vision.[3]

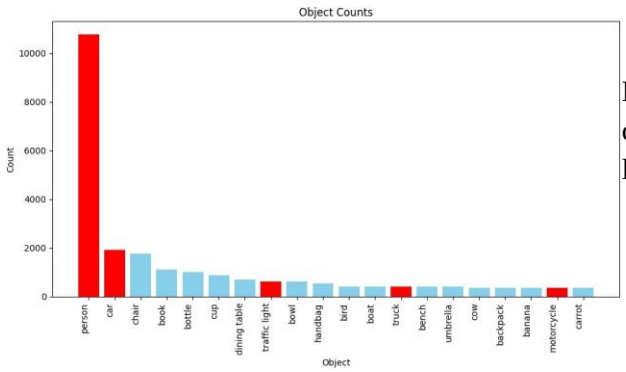

Figure 3.1. Class distribution in COCO (top 20 classes). The red bar represents the common labels

3.2. IDD

The IDD (Indian Driving Dataset) is a specialized repository meticulously crafted for research in autonomous driving and computer vision tasks, specifically tailored to the dynamic and varied driving conditions prevalent on Indian roads. Comprising annotated images and videos captured from dashcams and onboard sensors, the dataset encompasses diverse scenarios ranging from urban congestion to rural landscapes, with annotations including pixellevel semantic segmentation masks, bounding boxes, and metadata crucial for tasks like object detection and scene understanding.[1]

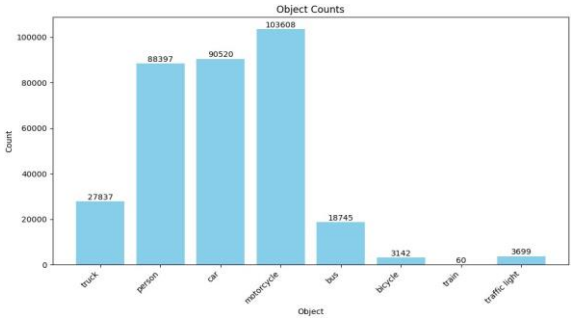

Figure 3.2. Class distribution in IDD (on 8 common labels)

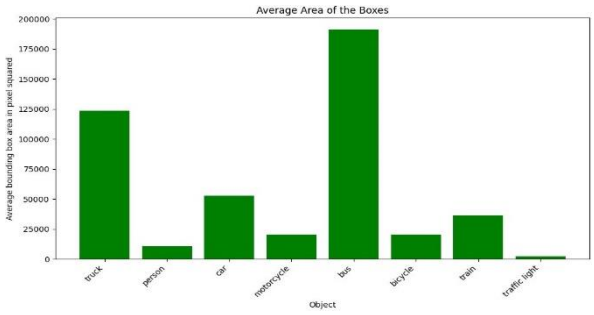

Figure 3.3. Average bounding box area of each class

4. Approach

We've successfully transformed both datasets into the required format for YOLOv8 and Vision Transformer objectdetection. This format comprises a directory housing an 'images' folder and a 'bounding box' folder. Each text filewithin the bounding box directory corresponds to an image in the images folder, mapping bounding boxes and their associated classes by filenames. Additionally, we've meticulously prepared a YAML file containing essential parameters for YOLOv8 and Vision Transformer object detection, including image paths, bounding box paths, and the numberof classes.

With the YAML file in place, we've executed the YOLOv8 and and Vision Transformer models to generate results and predictions. These outputs now serve as the basis for a rigorous analysis aimed at assessing both the successes and errors of the detection process.

5. Results

Upon conducting experiments, we observed that the reported results were successfully reproduced in our experimental setup. Specifically, models trained on the COCO dataset exhibited remarkable performance when evaluated on COCO's validation data, consistent with previous findings in the literature. However, when applied to the IDD dataset, these models yielded poor results. This discrepancy in performance may be attributed to the unique challenges present in the IDD dataset, such as occlusions and other traffic conditions that are characteristic of Indian roads. Theseconditions differ significantly from those encountered in the COCO dataset, highlighting the importance of dataset diversity and the need for specialized models to address specificenvironmental contexts.

5.1. COCO

SHODH SAGAR®
International Journal for Research Publication and Seminar
ISSN: 2278-6848 | Vol. 15 | Issue 3 | Jul - Sep 2024 | Peer Reviewed & Refereed

Classes	YOLO		RT-DETR	
	mAP50	mAP50-95	mAP50	mAP50-95
all	0.521	0.372	0.702	0.521
person	0.745	0.514	0.845	0.607
bicycle	0.456	0.264	0.646	0.396
car	0.561	0.363	0.735	0.498
motorcycle	0.652	0.412	0.803	0.56
bus	0.743	0.624	0.855	0.738
train	0.835	0.648	0.934	0.765
truck	0.437	0.296	0.643	0.458
traffic light	0.411	0.211	0.572	0.307

Table 1. Inference results on COCO

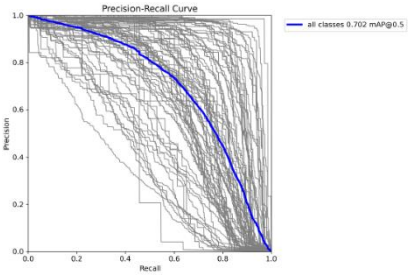

Figure 5.3. Precision Recall curve of RT-DETR

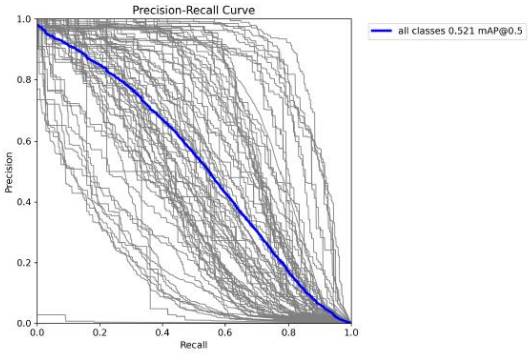

Figure 5.1. Precision-Recall curve of YOLOv8

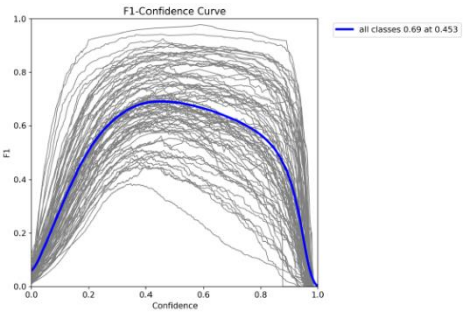

Figure 5.4. F1-Confidence curve of RT-DETR

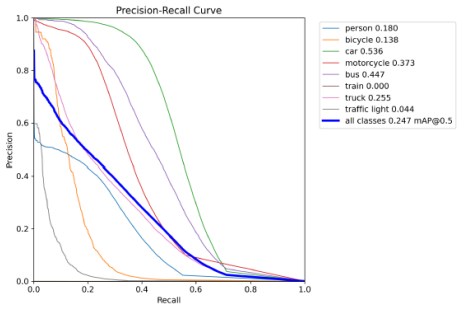

Figure 5.5. Precision-Recall curve of YOLOv8

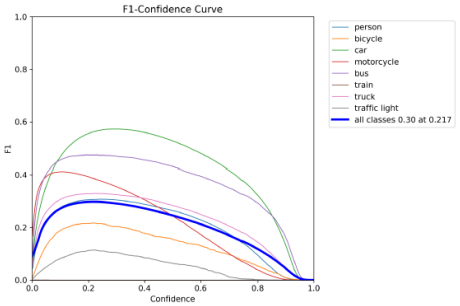

Figure 5.6. F1-Confidence curve of YOLOv8

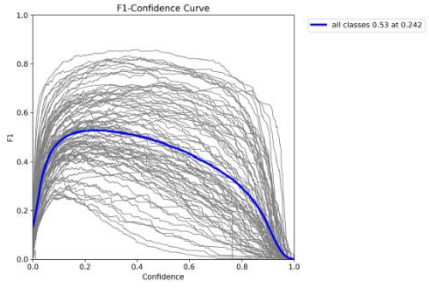

Figure 5.2. F1-Confidence curve of YOLOv8

5.2. IDD

Classes	YOLO		RT-DETR	
	mAP50	mAP50-95	mAP50	mAP50-95
all	0.247	0.161	0.355	0.228
person	0.18	0.101	0.24	0.136
bicycle	0.138	0.0788	0.301	0.167
car	0.536	0.371	0.67	0.468
motorcycle	0.373	0.182	0.523	0.254
bus	0.447	0.349	0.522	0.418
train	8.32e-05	3.97e-05	0.00427	0.0025
truck	0.255	0.185	0.379	0.282
traffic light	0.0437	0.0208	0.197	0.0987

Table 2. Inference results on IDD

SHODH SAGAR®
International Journal for Research Publication and Seminar
ISSN: 2278-6848 | Vol. 15 | Issue 3 | Jul - Sep 2024 | Peer Reviewed & Refereed

Experiments

Hypothesis 1

The experiment aimed to evaluate the performance of YOLOv8 in detecting small objects in both COCO and IDD datasets. Mean Average Precision, as implemented by Aladdin Persson, was adapted to extract ground truths predicted by the model and those that were not, regardless of classes or detection quality. All ground truths were normalized to a scale of 640 x 640, and an IoU threshold of 0.6 was used to classify a bounding box as a predicted or nonpredicted ground truth.

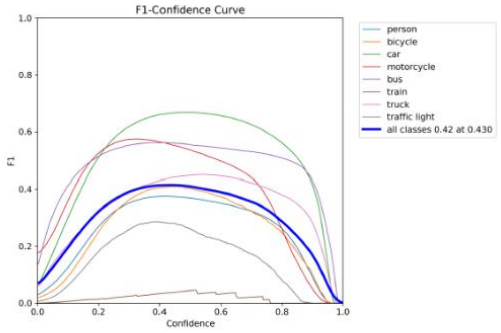

Figure 5.8. F1-Confidence curve of RT-DETR

Classes	COCO		IDD	
	notPred	Pred	notPred	Pred
Average Area	17512	48681	3121	21832
Count(1000 sq.units)[%]	38%	7%	72%	11%
Count (Total)	20024	16757	27429	12491

Table 3. Inference results on IDD

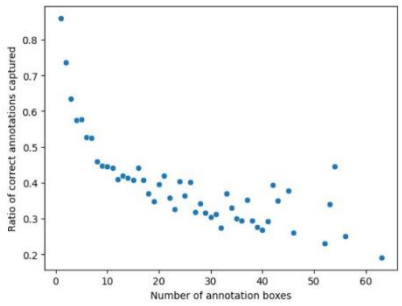

Figure 6.1. Ratio of correct annotations captured vs Number of annotation boxes for COCO

Results showed that for the COCO dataset, the average area of predicted ground truths was 48,681 sq. units, with 7% of them having an area less than 1000 sq. units. In contrast, the average area of non-predicted ground truths was 17,512 sq. units, with 38% of them having an area less than 1000 sq. units.

Similarly, for the IDD dataset, the average area of predicted ground truths was 21,832 sq. units, with 11% of them having an area less than 1000 sq. units. The average area ofnon-predicted ground truths was 3,121 sq. units, with 72%of them having an area less than 1000 sq. units.

These findings suggest that YOLOv8 struggles with detecting small objects, which contributes to model error. Possible reasons for this difficulty could include limitations infeature representation or the anchor box configuration usedin the YOLOv8 architecture.

5.2. Hypothesis 2

For the COCO dataset, a thorough examination was conducted to assess the model's performance in object detection. Among the 4900 images, it was discovered that in a significant portion, approximately 1850 images, the modelfailed to recognize ground truths. Specifically, these imagespresented a scenario where over 50% of the ground truths remained undetected, even when employing a stringent 0.6 Intersection over Union (IoU) threshold.This observation highlights a considerable challenge faced by the model in accurately identifying objects within the COCO dataset. Tofurther explore the extent of this challenge, a closer examination was conducted on the subset of images with the highest number of ground truths. This analysis aimed to highlight whether the model's performance varied significantlybased on the density of objects within an image. Through graphical representation, it was revealed how the percentage of undetected ground truths fluctuated across the top 100 images with the most ground truths, shedding light on potential patterns or anomalies in the model's behavior under varying object densities.

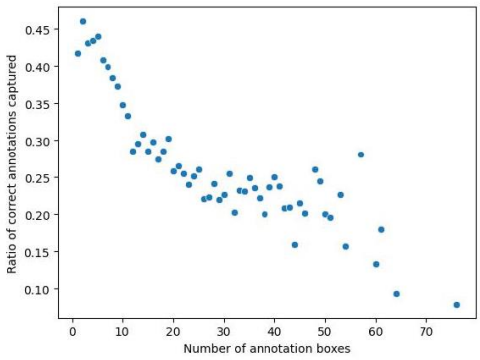

Figure 6.2. Ratio of correct annotations captured vs Number of annotation boxes for IDD

Similarly, the investigation extended to the IDD dataset,which encompasses a diverse array of urban scenes captured from onboard vehicle cameras. Among the 4762 images scrutinized from this dataset, a noteworthy trend emerged, with 3078 images exhibiting a significant shortfall in ground truth detection. Once again, employing the
0.6 IoU threshold criterion, more than 50% of ground truthsremained undetected in these images, indicative of themodel's challenges in accurately identifying objects withinurban environments. In essence, the findings from both the COCO and IDD datasets underscore the nuanced challenges encountered by the model in object detection tasks, ranging from diverse object categories to varying environmental contexts. By meticulously analyzing the prevalence of undetected ground truths across a substantial number of images, this experiment provides valuable insights into thelimitations and areas for improvement in contemporary object detection models.

References

[1] Indian driving dataset. https://idd.insaan.iiit.ac.in/. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, NicolasUsunier, Alexander Kirillov, and Sergey Zagoruyko. Detr: End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*,2020. 1

[3] Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C. Lawrence. Microsoft COCO: Common Objects in Context. http://cocodataset. org/, 2014. 2

[4] Aladdin Persson. Machine learning collection.

[5] Ultralytics. YOLOv8 GitHub Repository. https://github.com/ultralytics/yolov8, 2022. 1