

Enhancing Object Detection with YOLOv8 Transfer Learning: A VOC2012 Dataset Study

Nan Zhao ^a

Computer Science, University of Liverpool, Liverpool, U.K.

Keywords: Object Detection, Transfer Learning, YOLOv8, VOC2012 Dataset.


Abstract: Object detection has received widespread attention due to its use in a large number of practical scenarios. In this paper, the primary objective is to develop and evaluate a transfer learning-based object detection framework using the you only look once v8 (YOLOv8) model. The study investigates the performance and parameter influence of YOLOv8 when trained on custom datasets through transfer learning methodologies. Firstly, this paper introduces the Vision Object Classes (VOC) 2012 dataset as the primary input data. Subsequently, the YOLOv8 model is configured as the foundational network for feature extraction and object detection tasks. Additionally, transfer learning techniques are applied to enhance the model's generalizability. Thirdly, key parameters are adjusted and compared for thorough analysis. Furthermore, the YOLOv8 predictive performance is assessed and juxtaposed with results obtained from the COCO dataset. The experimental findings highlight the robust performance of YOLOv8 on the VOC2012 dataset. This study offers valuable insights and serves as a reference for researchers in the field, shedding light on effective strategies for object detection using transfer learning approaches.

1 INTRODUCTION

Object detection is an important task in the computer vision, which is one of core problems in computer vision (Ma, 2024). The task of object detection finds objects that all of objects, images and videos are interested in, and makes sure classification and position of object. Light interference and occlusion in target detection processes is challenging on computer vision because different objects have different appearance, posture, imaging (Ma, 2024). Since the introduction of deep learning with object detection, most of object detectors are divided into two classes. Firstly, detection methods are region-based. System based on region is known two-stage detectors, which makes sure the candidate box of the sample and then using convolutional neural networks (CNN) (such as Regional CNN(R-CNN)) classifies the samples (Ma, 2024). Secondly, detection methods are regression-based. System based on regression is known single-stage detectors, which does not product the candidate box of the sample in processing and implement target detection directly based on specific regression (Ma,

2024). The difference of two classes is that detector accuracy of two steps is better than one step detector, but real-time detectors performance of two steps is low (Ma, 2024). Researchers pay more attention to the each-step detector because each-step detector has comprehensive performance and excellent operating efficiency, you only look once (YOLO) is the most common example of the single step detector (Ma, 2024).

YOLO models mean “You Only Look Once” series of models about object detect, which have very high influence on computer vision (Ma, 2024). YOLO models keep real-time performance and achieves high accuracy, which apply for surveillance, vehicle navigation and automatic image labelling, etc (Ma, 2024). YOLOv8 is one model of “You Only Look Once” series of object detection models, which is famous for efficiency and speed in the detecting objects of images and videos (Jönsson, 2023). There are some key features and enhancements of YOLOv8. YOLOv8 continually improve detection accuracy in order to become one of the fastest and the most accurate object detect models (Soylu, 2023).

^a <https://orcid.org/0009-0003-3262-8628>

Architecture of YOLOv8 is refined and optimized, which increases performance on a variety of hardware platforms from high-end GPUs to edge devices and makes sure wider accessibility and applicability (Xiao, 2023). YOLOv8 includes the latest advances in machine learning and deep learning and uses enhanced training techniques to improve model ability from generalize training data to real-world scenarios, which decreases overfitting and improve detection performance on unseen data (Das, 2024). YOLOv8 is more robust to different types of input data differences (such as lighting, occlusions, and objects of different scales), which makes more reliable in the diverse and challenging environment (Aboah, 2023). YOLOv8 usually supports much wider object class, which makes more versatile and useful in the different fields and applications (Motwani, 2023). YOLOv8 is easily integrated with existing software and platforms and has more compatibility and support for various programming languages and frameworks (Luo, 2024). Although YOLOv8 represents the cutting edge of object detection technology, each iteration of YOLOv8 focuses on balancing trade-offs between speed, accuracy and computational requirements (Reis, 2023).

The main object of this research analyses a transfer learning that bases on object detection framework using the YOLOv8 model. The research investigates the performance and parameter influence of YOLOv8 when trained on custom datasets through transfer learning methodologies. Firstly, the paper introduces the VOC2012 dataset as the primary input data. Subsequently, the YOLOv8 model is configured as the foundational network for feature extraction and object detection tasks. Additionally, transfer learning techniques are applied to enhance the model's generalizability. Thirdly, key parameters are adjusted and compared for thorough analysis. Furthermore, the YOLOv8 predictive performance is assessed and juxtaposed with results obtained from the COCO dataset. The experimental findings highlight the robust performance of YOLOv8 on the Vision Object Classes (VOC) 2012 dataset. This study offers valuable insights and serves as a reference for researchers in the field, shedding light on effective strategies for object detection using transfer learning approaches.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

The PASCAL VOC dataset (Pascal, 2012) is widely used for object detection, segmentation, and classification tasks, serving as a benchmark for computer vision models. It comprises two main challenges, VOC2007 and VOC2012, with annotations including bounding boxes and class labels for object detection and segmentation masks. The VOC2012 subset, used in this study, contains 11,530 images with annotations for 27,450 objects and 6,929 segmentations. Data preprocessing involves resizing and normalization, with a Python script converting annotations into the required format for YOLOv8. This project employs the VOC dataset to train and analyze YOLOv8's performance, comparing it against the original COCO dataset.

2.2 Proposed Approach

The main purposes of this paper are to construct and analysis target detection based on transfer learning by using YOLOv8 model (as shown Figure 1). Using transfer learning research YOLOv8 training on performance of custom dataset and influence of parameters. Specifically, firstly the VOC2012 dataset is used to the data input for this paper. Secondly, Constructing the YOLOv8 model is as the basic network for feature extraction and target detection tasks, at the same time, using transfer learning techniques generalizes model performance further. Thirdly, core parameters need to be adjusted and comparatively analysed. Additionally, this paper analysis performance of prediction based on YOLOv8 model, and compares with the COCO dataset.

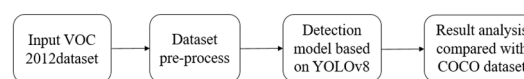


Figure 1: The framework of the YOLOv8 model (Photo/Picture credit: Original).

2.2.1 YOLOv8

YOLOv8 mainly uses the Ultralytics framework and the core of feature shows in the Figure 2, YOLOv8 provides a new the model of SOTA that includes P5 640 and P6 1280 resolution of object detect network and model of example segmentation by using YOLACT. Factors of scaling provide models of

different sizes in N/S/M/L/X scales in order to satisfy different requirements, which is the same as YOLOv5. The main YOLOv8 model has network of a backbone, head of a detection, and function of a Loss. Typically, the layer of input is an image of predefined size, which often is a mage size multiple of 32.

The network of backbone and Neck refer to the YOLOv7 design idea, using the C2f (CSPLayer_2Conv) structure with richer gradient flow replaces the C3 structure of YOLOv5, and adjusting different channel numbers for different scale models.

The model structure of YOLOv8 is fine-tuned and greatly improves the performance of model. But YOLOv8 model exists some operations such as Split are, which is not as friendly to specific hardware deployments as before (Reis, 2023).

The Head part has more changing than YOLOv5, which replaces by the current mainstream decoupled head structure, and separates the classification and detection heads, Anchor-Based changes Anchor-Free at the same time (Reis, 2023). YOLOv8 uses dual FPN that is similar with feature extracting mechanism of PAN-FPN on the YOLOv5 in the feature detector. Dual-FPN extracts uses multiple convolution and pooling operations to effectively extract information of features that are from the input image, which quickly and effectively achieves performance of detection (Reis, 2023).

The detection layer is responsible to task of object detection that is features extracted through the network of backbone, YOLOv8 predicts bounding box locations and classes through specific convolutional and connection layers. The output layer is used to output information of detection object that includes its category, location, and score of confidence. Result of detection object by the output. The final output need use Non Maximum Suppression (NMS) to eliminate bound boxes of overlap, and use highest confidence as box of the bounding (Reis, 2023). The calculation of loss uses positive and negative sample allocation strategy of TaskAlignedAssigner. The training data augment need close enhancement of mosaic in YOLOX for the final10 epochs, which could effectively increase the accuracy (Reis, 2023). In this paper, using YOLOv8 model trains VOC2012 dataset and COCO dataset in order to analysis performance of objection detection.

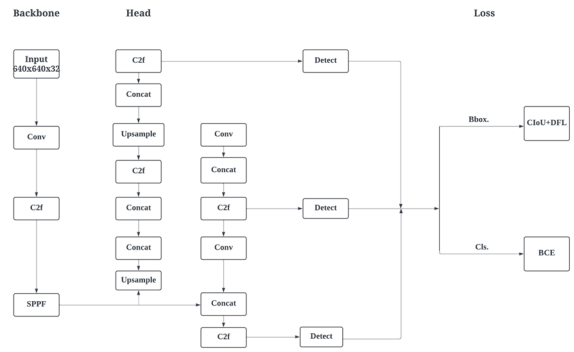


Figure 2: Model structure of YOLOv8 detection model (Photo/Picture credit: Original).

2.2.2 Loss Function

The process of Loss calculation has two parts that are positive and negative sample allocation strategy and Loss calculation. The YOLOv8 model firsthand uses TaskAlignedAssigner of TOOD due to strategy excellence of the dynamic allocation. TaskAlignedAssigner mainly uses classification and regression scores that are weighted to select positive samples.

$$t = s^\alpha + \mu^\beta \tag{1}$$

s is score of the prediction about the annotation classes, u means the Intersection Over Union (IoU) that is the prediction box and the Ground Truth (GT) box, and multiplying of s and u is the degree of alignment.

For each GT, all prediction boxes are based on the classification score about the GT classes. The weighting of IoU that is the prediction box and GT obtains alignment metrics of an alignment score related to classification and regression. The maximum top K is directly selected based on the alignment metrics alignment score as a positive sample. Loss calculation includes classification and regression branches without the previous objection branch. Binary cross entropy (BCE) Loss is used to the branch of classification. Distribution Focal Loss and IoU Loss is used to the regression branch. Three Losses require a certain ratio of weight plus.

2.3 Implementation Details

This paper chooses to pre-training to initialize the model and then training on a VOC2012 dataset and COCO dataset. These weights are from a model trained on the VOC2012 dataset and COCO dataset. Hyper-parameter tuning approach is used by having access to RTX 3080Ti. Configuration file uses

YAML and modifying hyper-parameters uses CLI parameters. Using VOC2012 dataset that the image size is 640 trains YOLOv8n for 500 epochs. Using two GPUs that CUDA devices are 2 and 3 trains YOLOv8n model. Using COCO dataset that the image size is 640 trains YOLOv8n for 500 epochs. Using three GPUs that CUDA devices are 0, 1, 2 trains YOLOv8n model, because multi-GPU training makes more efficient use of available hardware resources by distributing training load across multiple GPUs.

In this paper, box loss of training and class loss of training, distribution focal loss of training respectively present boxes loss of the bound in the train, class loss of training, and distribution focal loss of training. Validation box loss and validation class loss loss, validation distribution focal loss respectively indicates boxes loss of the bound in the validation, class loss of validation, and distribution focal loss of validation. Mean average precision (mAP) is a popular evaluation metric in the object detection, mAP uses average precision of all classes and maculates them according to pre-specified threshold of intersection over union. mAP50(B) of metrics and recall(B) of metrics use the metrics of model evaluation, intersection of union (IoU) in the mean average precision(mAP) is 0.50 that means the mAP50, B is a type of segmentation metrics, for example, precision (B) of metrics is detection and precision(M) of metrics is segmentation (Reis, 2023). Metrics/mAP50 (B) indicates that using steps of 0.05 starts from an IoU threshold of 0.5 and stops at 0.95, taking average precision (AP) in this interval then all of classes do this and take average in all of classes, which product mAP50 (B) of Metrics (Reis, 2023). Precision-recall visualizes the balance in precision and recall, which make accuracy selective prioritize, the chosen threshold determines the specific precision-recall values (Reis, 2023). F1-confidence is a measure of balanced performance at different confidence levels, which combines precision and recall, the peak means the confidence level where the model arrives the best trade-off both precision and recall in the specific dataset (Reis, 2023).

3 RESULTS AND DISCUSSION

This research compares VOC2012 dataset with COCO dataset about performance of object detection by training YOLOv8 model. Using the train box loss, train class loss, train distribution focal loss (dfl), precision(B) of metrics, recall(B) of metrics, mAP50(B) of metrics, mAP50-95(B) of metrics,

validation box loss, validation class loss, distribution focal loss of validation, precision and recall, F1-Confidence and Precision-Recall to evaluate performance of object detection.

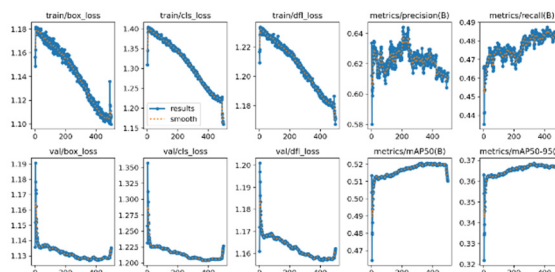


Figure 3: The result of COCO dataset (Photo/Picture credit: Original).

Figure 3 shows that box loss (box_loss), distribution focal loss (dfl_loss) and class loss (cls_loss) determine the best epochs (490) on the best result of object detect about coordinates of bounding box in the train and validation COCO dataset. Precision (B) has the selection of the best epochs (240) and Recall (B) has the selection of the best epochs (321) during training. The Map50 (B) and Map50-95 (B) have the selection of the best epochs (391). Analysis of these curves enables the selection of the best epochs in order to achieve best results in object detection on the COCO dataset.

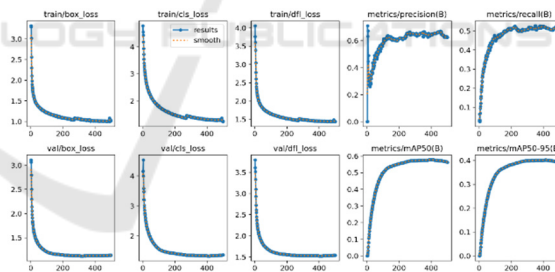


Figure 4: The result of VOC2012 dataset (Photo/Picture credit: Original).

Figure 4 shows that box loss (box_loss), distribution focal loss (dfl_loss) and class loss (cls_loss) determine the best epochs (400) for the best object detection about the coordinates of bounding box during train and validation VOC2012 dataset. Precision (B) has the optimal number of epochs (161) and Recall (B) has the best epochs (223) during training. The Map50 (B) has the best epochs (161) and Map50-95 (B) have the optimal number of epochs (397). Analysis of these curves enables the selection of the best epochs (240) in order to achieve the best results in object detection on the VOC2012 dataset.

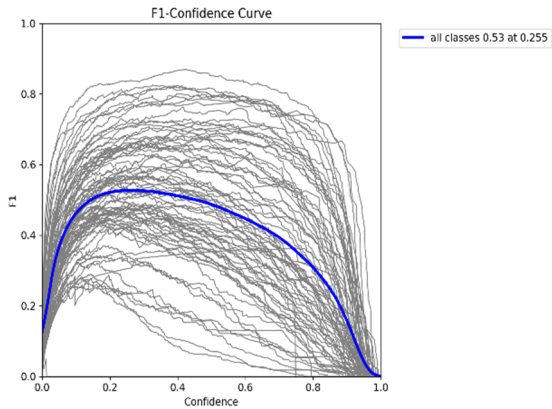


Figure 5: F1-Confidence on the COCO dataset (Photo/Picture credit: Original).

"All classes 0.53 mAP at 0.255" means that YOLOv8 model of the object detection achieved an average precision of 53% across all classes of objects, with a confidence threshold of 0.255 for determining correct detections on the COCO dataset (as shown Figure 5).

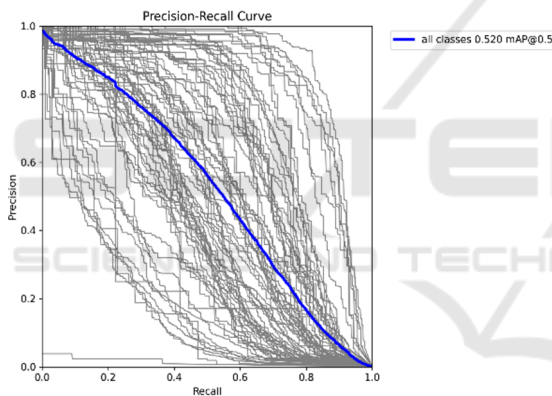


Figure 6: Precision-Recall on COCO dataset (Photo/Picture credit: Original).

"All classes 0.52 mAP @ 0.5" indicates that the YOLOv8 model of object detection gets a mean Average Precision that is 0.52 through all classes in the dataset, evaluation of predicted accuracy on the bounding boxes in the COCO dataset uses an Intersection of Union threshold that is 0.5 (as shown Figure 6).

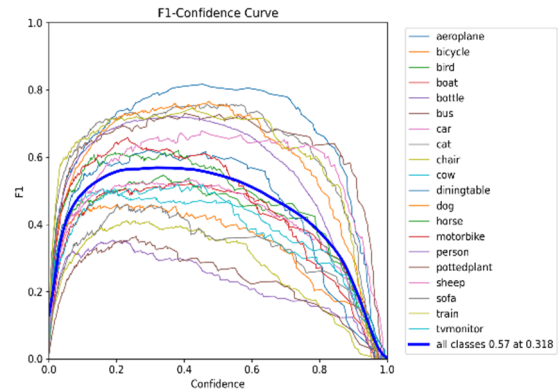


Figure 7: F1-Confidence on the VOC2012 dataset (Photo/Picture credit: Original).

"All classes 0.57 mAP at 0.318" means that YOLOv8 model of the object detection achieved an average precision of 57% across all classes of objects, with a confidence threshold of 0.318 for determining correct detections on the COCO dataset (as shown Figure 7).

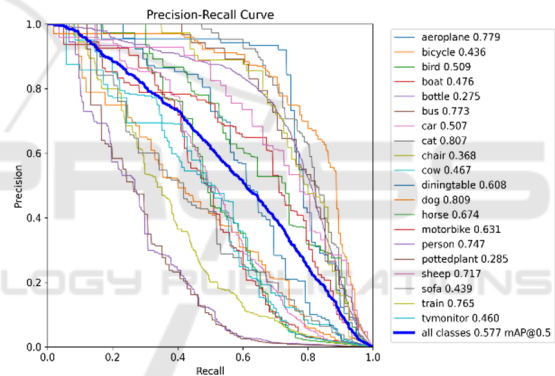


Figure 8: Precision-Recall on VOC2012 dataset (Photo/Picture credit: Original).

"All classes 0.577 mAP @ 0.5" indicates that the YOLOv8 model of object detection gets a mean Average Precision that is 0.577 through all classes in the dataset, evaluation of predicted accuracy on the bounding boxes in the COCO dataset uses an Intersection of Union threshold that is 0.5 (as shown Figure 8).

Table 1: Comparing performance in COCO and VOC2012 dataset on the YOLOv8.

dataset	F1- confidence	Precision- Recall
COCO	0.53 mAP at 0.255	0.52 mAP @ 0.5
VOC2012	0.57 mAP at 0.318	0.577 mAP @ 0.5

F1-confidence is 0.57 mAP at 0.318 and Precision-Recall is 0.577 mAP @0.5 on the VOC2012 dataset, as shown in the Table 1. F1-confidence is 0.53 mAP

at 0.255 and Precision- Recall is 0.52 mAP @0.5 on the COCO dataset. F1-confidence and Precision-Recall on the VOC2012 dataset are higher than COCO dataset VOC2012 from the table1. A higher mAP score generally indicates better performance. Therefore, the performance of VOC2012 dataset on the YOLOv8 is better than VOC2012 dataset on the YOLOv8.

4 CONCLUSIONS

This study introduces an investigation into object detection using the YOLOv8 model applied to both the COCO and VOC2012 datasets. The primary objective is to assess the performance of the model by analyzing key metrics such as F1-Confidence and Precision-Recall across these datasets. Upon comparison, it is evident that the VOC2012 dataset outperforms the COCO dataset in terms of F1-Confidence and Precision-Recall scores. Through meticulous experimentation and analysis, the study confirms the superior performance of the YOLOv8 model on the VOC2012 dataset relative to COCO. Looking ahead, future research endeavors will prioritize expanding the dataset size to enhance the robustness and accuracy of the analysis. It is worth noting that the disparity in dataset sizes, with VOC2012 being smaller than COCO, may have implications on the obtained results, indicating the importance of dataset selection and size in object detection studies.

REFERENCES

- Aboah, A., Wang, B., Bagci, U., & Adu-Gyamfi, Y. 2023. Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5349-5357.
- Das, B., & Agrawal, P. 2024. Object Detection for Self-Driving Car in Complex Traffic Scenarios. In MATEC Web of Conferences. vol. 393, p: 04002.
- Jönsson Hyberg, J., & Sjöberg, A. 2023. Investigation regarding the Performance of YOLOv8 in Pedestrian Detection.
- Luo, X., Zhu, H., & Zhang, Z. 2024. IR-YOLO: Real-Time Infrared Vehicle and Pedestrian Detection. Computers, Materials & Continua, vol. 78(2).
- Ma, S., Lu, H., Liu, J., Zhu, Y., & Sang, P. 2024. LAYN: Lightweight Multi-Scale Attention YOLOv8 Network for Small Object Detection. IEEE Access.
- Motwani, N. P., & Soumya, S. 2023. Human Activities Detection using DeepLearning Technique-YOLOv8. In ITM Web of Conferences. vol. 56, p: 03003.
- Pascal, V., 2012. VOC dataset. <https://docs.ultralytics.com/datasets/detect/voc/>
- Reis, D., Kupec, J., Hong, J., & Daoudi, A. 2023. Real-time flying object detection with YOLOv8. arXiv:2305.09972.
- Soylu, E., & Soylu, T. 2023. A performance comparison of YOLOv8 models for traffic sign detection in the Robotaxi-full scale autonomous vehicle competition. Multimedia Tools and Applications, pp: 1-31.
- Xiao, X., & Feng, X. 2023. Multi-object pedestrian tracking using improved YOLOv8 and OC-SORT. Sensors, vol. 23(20), p: 8439.