University of New South Wales

COMP9444

Neural Networks and Deep Learning

Project Report

Pedestrian detection of crowded pedestrian images
Group name: Random

Ebubekir Stuart Clark	z5258638
	z5258638@ad.unsw.edu.au
Lin Zhang	z5386409
	z5386409@ad.unsw.edu.au
Yu Tong	z5306062
	z5306062@ad.unsw.edu.au
Hao Wang	z5346755
	z5346755@ad.unsw.edu.au
NAME	ZID
	zxxxxxxx@ad.unsw.edu.au

Submitted on August 4, 2023

Contents

1 Overview

Deep learning architecture is commonly used for classification of pedestrians for applications such as in autonomous vehicles. However, these models perform less accurately for crowded pedestrian images Gao et al. 2023. Here we apply modifications of state of the art neural network architecture for applications in classification of crowded pedestrian settings, inspired from current research. We base our architecture on the *You Only Look Once* abbreviated *YOLO* version 5 architecture, herein referenced as YOLOv5n (see Jocher 2020), a popular network architecture for pedestrian classification due to prioritisation of quick classification. We train and test this data on the benchmark crowded person dataset CrowdHuman from Shao et al. 2018, which contains 15000, 4370, 5000, training, validation, and testing images respectively. We use a modified Non-Maximum Supression (NMS) region using a distance based intersection over union region (DIoU) as suggested by Zheng et al. 2019 and compare performance on crowded images using a regular NMS algorithm using IoU regions.

2 Literature Review

Gao et al. 2023 states that crowded pedestrian detection in object detection remains a challenging problem. This paper suggests some improvements to the YOLO architecture to improve the Intersection over Union (IoU) loss.
Several other improvements for image classification have been suggested by researchers. Huang et al. 2020 suggests that using a Non-Maximum Suppression (NMS) leads to missing highly overlapped pedestrians. The authors suggest using a modified NMS region hereby called Representitive Region NMS (RRNMS) to improve classification performance, benchmarked against the CrowdHuman and CityPersons datasets.
In a similar light, Zheng et al. 2019 proposes an improvement to the classic IoU distance metric, using a distance based approach (DIoU). It seems that the variations of NMS regions have great significance in the prediction of pedestrians in crowded scenes, hence we hypothesise that using this DIoU metric may improve the performance of crowded pedestrian detection.

3 Methodology

In this study, we present a novel methodology to enhance the object detection performance in YOLOv5 by replacing the conventional Non-Maximum Suppression (NMS) technique with a custom implementation based on the Distance Intersection over Union (DIoU) metric. The motivation behind this approach is to address the limitations of traditional NMS, which tends to favor boxes with high IoU values without considering their spatial relationships or localization accuracy.
To achieve this, we introduced a custom DIoU calculation function, "bbox_iou", which accurately computes the DIoU between a single bounding box and multiple bounding boxes. DIoU takes into account the distance between bounding box centers, encouraging tighter bounding boxes that better align with the actual object boundaries. This results in improved localization precision and reduced localization errors, leading to more accurate and reliable object detection.
Subsequently, we developed a new NMS function, nms, utilizing the DIoU metric for box comparison and selection in "utils/general.py". The nms function takes bounding boxes, corresponding confidence scores, and a user-defined DIoU threshold as inputs. During the NMS process, the DIoU threshold serves as a critical parameter to control the degree of overlap allowed between bounding boxes. By setting an appropriate DIoU threshold, we can strike a balance between suppressing redundant detections and preserving potentially meaningful but spatially overlapping objects. This fine-tuning capability allows us to achieve optimized performance in different detection scenarios.
We implement this by replacing the default torch.ops.nms with the a custom nms function, which can be found in utils/general.py in the project directory.

We use the baseline architecture YOLOv5n, since it has the smallest number of parameters (1.9 million) of the YOLO models and can be trained relatively quickly, details can be found in Jocher 2020. After training with image dimensions 608×608, and batch size of 16 (with the rest as default YOLOv5 parameters), we apply the previously mentioned DIoU algorithm to the boxes returned by the model with the aim of improving detection of crowded places.

4 Results

After training the model for 6 epochs with image size being 608×608 dimensional, a batch size of 16, and the rest of the hyperparameters as defaults for YOLOv5, we receive the following results summarised in table 1 below.

	Precision	Recall
IoU Loss Function	0.74336	0.52677
DIoU Loss function	0.6749	0.34062

Table 1: Precision and recall metrics using IoU loss and DIoU loss

Unfortunately, the DIoU loss function does not achieve the same precision and recall as the IoU loss function for the CrowdHuman dataset using the parameters specified earlier, despite being more computationally demanding. We justify that 6 epochs were sufficient for training given that the precision was converging, as can be demonstrated below in figure 1 using the IoU loss convergence.

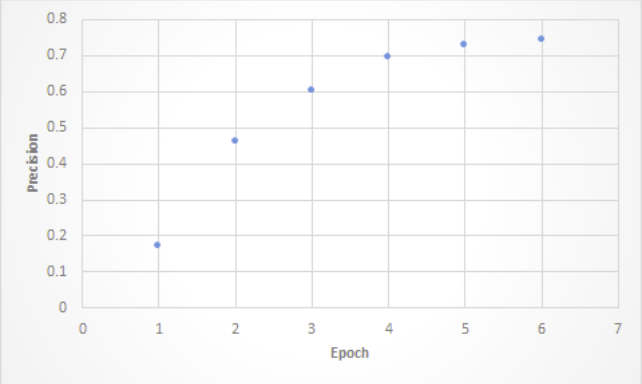

Figure 1: Precision of the IoU loss variant per epoch

The images below are examples of classification applied to test set images, demonstrating before and after classification in figures 2 and 3 respectively.

Figure 2: Test set example

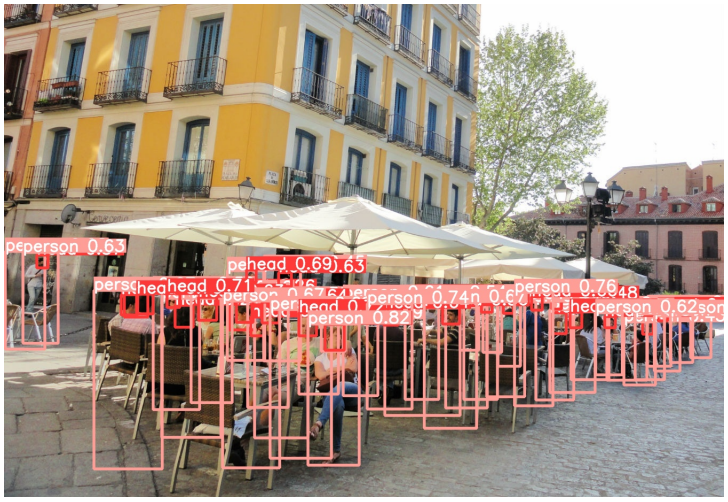

Figure 3: Classified test set example

Above we see that regardless of its limitations, the bounding boxes for the two classes *person* and *head* look impressive, detecting even small faces in the distance amongst many other people.

5 Conclusion

The DIoU-based Non-Maximum Suppression (NMS) approach should effectively optimizes pedestrian detection in crowded scenes by efficiently suppressing redundant detections, ensuring that only the most relevant bounding boxes are retained. This reduction in false positives significantly should enhance overall detection accuracy, a critical factor in densely populated environments. Furthermore, the focus on precise localization enables the model to accurately identify pedestrian boundaries, adding particular value to tasks like instance segmentation and fine-grained object detection.
Nevertheless, it is important to consider some limitations introduced by the DIoU-based approach. The method involves complex geometric calculations, resulting in higher computation complexity compared to traditional IoU-based NMS. This might lead to increased inference time, particularly in scenarios with large-scale datasets or dense object arrangements, necessitating a careful trade-off between computational cost and potential accuracy gains, especially in

resource-constrained settings. However, we aimed to offset this limitation by using the smallest architecture for YOLO (YOLOv5n) in the number of parameters.
Additionally, selecting an appropriate DIoU threshold is pivotal for optimal detection results, requiring fine-tuning according to specific tasks and datasets. Overly strict thresholds may exclude valid detections, whereas lenient thresholds could introduce false positives. Hence, hyperparameter optimization is crucial to maximizing the DIoU-based NMS performance.
However, the results do not support this hypothesis. This could be due to the fact that hyperparameter tuning was not conduced for the DIoU threshold due to immense computational constraints. Unfortunately, whilst the YOLOv5n model is relatively simple, it did not offset the computational intensity significantly to train the model with the DIoU loss function. This limitation is partially due to hardware constraints, but also must be attributed to an increased computational overhead.
In the future we would repeat this experiment by using better hardware and hyperparameter optimisation techniques, such as grid search, to improve the performance of the model using DIoU, to fairly compare the two performances.

6 Optimisation

As part of this project, we also explored optimisation techniques that could improve the computational and quantitative performance of YOLO for crowded pedestrian detection. *[1911.11907] GhostNet: More Features from Cheap Operations* 2023 suggested that using a novel Ghost module to generate more feature maps from cheap operations would yield in faster and more accurate classification of yolo architecture. Due to time and hardware constrains we were unable to train our DIoU model with these changes, however these modifications were tested and implemented. In short, the changes to the YOLOv5 structure are presented as follows:

1. Replacing the traditional convolution, BatchNormalization, and LeakyRelu modules in the backbone section with GhostBottleNeck
2. Using TCSP Ghost2 replaces the C3 module in the backbone section
3. Replacing the C3 and Conv module in the head section with GSCONV and VoVGSCSP

Our results demonstrated nearly a 50% reduction in file size of the weight file, allowing for a more lightweight model, while retaining the performance of the original classifier as demonstrated below in figure 4.

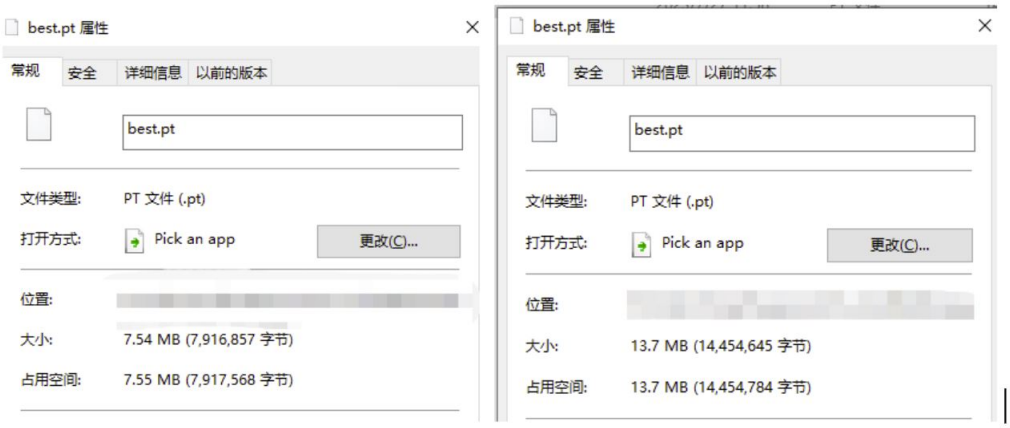

Figure 4: Trained weight file sizes. GhostNet backbone (left), Default (Right),

In the future we would wish to integrate this work into the DIoU trained model to improve its performance, however, this is outside the scope of this project.

References

[1911.11907] GhostNet: More Features from Cheap Operations (2023). URL: https://arxiv.org/abs/1911.11907 (visited on 08/04/2023).

Gao, Fei et al. (Feb. 2023). "Improved YOLOX for pedestrian detection in crowded scenes". en. In: *Journal of Real-Time Image Processing* 20.2, p. 24. ISSN: 1861-8219. DOI: 10.1007/s11554-023-01287-7. URL: https://doi.org/10.1007/s11554-023-01287-7 (visited on 08/01/2023).

Huang, Xin et al. (2020). *NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing*. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Huang_NMS_by_Representative_Region_Towards_Crowded_Pedestrian_Detection_by_Proposal_CVPR_2020_paper.html (visited on 07/20/2023).

Jocher, Glenn (2020). *YOLOv5 by Ultralytics*. Version 7.0. DOI: 10.5281/zenodo.3908559. URL: https://github.com/ultralytics/yolov5.

Shao, Shuai et al. (2018). "CrowdHuman: A Benchmark for Detecting Human in a Crowd". In: *arXiv preprint arXiv:1805.00123*.

Zheng, Zhaohui et al. (Nov. 2019). *Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression*. en. URL: https://arxiv.org/abs/1911.08287v1 (visited on 08/01/2023).