Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230208

Exploration and evaluation of faster R-CNN-based pedestrian detection techniques

Shengxin Gao

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, 710049, China

2206113784@stu.xjtu.edu.cn

Abstract. Presented herein is an exploration into the efficacy of a pedestrian detection model that capitalizes on the Faster region-based convolutional neural network (R-CNN) algorithm. This model, following its training phase on the Caltech Pedestrian dataset, underwent a meticulous evaluation process designed to gauge its proficiency in pedestrian detection tasks. The Average Precision (AP) achieved during testing was an impressive 51.9%, pointing to a high degree of accuracy. In addition to its commendable accuracy, this model demonstrates a remarkable speed of inference. Each image was processed in a mere 0.07 seconds, underlining the model's potential for real-time pedestrian detection in real-world scenarios. Further enhancing its potential for deployment is the model's relatively compact storage footprint, consuming only 158MB of storage space. By providing an in-depth analysis of this Faster R-CNN-based pedestrian detection model, the study offers valuable insights for future developments in the computer vision field, particularly for real-world applications. This paper thus contributes to our understanding of the applicability and effectiveness of the Faster R-CNN algorithm, providing a solid foundation for future research and development.

Keywords: pedestrian detection, faster R-CNN, Caltech Pedestrian.

1. Introduction

Pedestrian detection stands as a crucial task within the realm of computer vision, with applications spanning autonomous driving, surveillance systems, and human-computer interaction. The accurate detection of pedestrians in images and videos presents a formidable challenge due to factors such as occlusion, complex backgrounds, and others. To address these challenges, myriad object detection algorithms have been developed over time.

The focus of this paper lies in constructing and evaluating the performance of a prominent object detection algorithm, the Faster R-CNN, specifically within the scope of pedestrian detection. The evaluation takes into consideration the widely recognized Caltech Pedestrian Dataset, a diverse compilation of pedestrian images that acts as a benchmark for gauging the effectiveness of pedestrian detection algorithms. Historically, pedestrian detection research has seen substantial advancements with the development of a variety of techniques and algorithms. For instance, extensive exploration into effective feature representations, such as HOG and its variants, has been conducted, aiming to capture local shape information. Additionally, the advent of deep learning has contributed significantly to pedestrian detection, with proposed CNN-based architectures like the Pedestrian Alignment Network

DOI: 10.54254/2755-2721/32/20230208

(PAN) and Bi-box Regression Network (Bi-box) incorporating body part alignment and scale variation handling respectively. In recent years, Faster R-CNN has garnered commendation for its impressive performance in object detection tasks. The algorithm integrates a Region Proposal Network (RPN) to generate candidate object regions, followed by a region-based convolutional neural network for accurate localization and classification. The multi-stage architecture of Faster R-CNN enables iterative refinement of detections, thereby leading to an enhancement in detection accuracy. As shown in Figure 1.

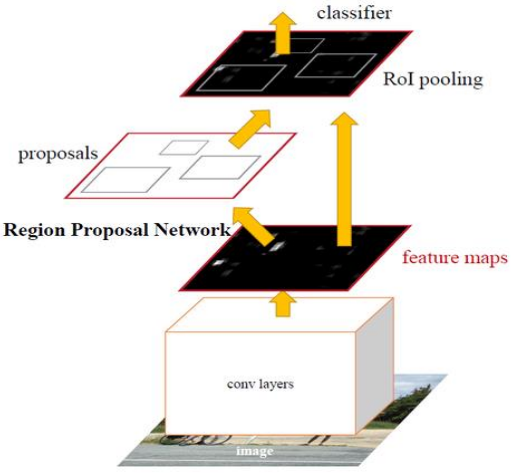


Figure 1. The overview process of faster R-CNN (Photo/Picture credit: Original).

The Caltech Pedestrian Dataset has emerged as a widely adopted benchmark for evaluating pedestrian detection algorithms. It comprises a large collection of frames captured from various surveillance cameras, providing detailed annotations of pedestrians with bounding box coordinates and occlusion labels. This extensive dataset facilitates comprehensive evaluation of detection algorithms. By evaluating the Faster R-CNN algorithm on the Caltech Pedestrian Dataset, this study aims to provide insights into the strengths and weaknesses of this algorithm for pedestrian detection. Performance evaluation will be conducted using standard metrics, including average precision, inference time, resource requirements and so on, considering varying levels of occlusion and scale variations. The findings of this study will contribute to a deeper understanding of Faster R-CNN object detectors in detecting pedestrians. By evaluating the performance of the algorithm on the Caltech Pedestrian Dataset, valuable insights can be gained regarding the effectiveness and applicability of this algorithm in real-world pedestrian detection scenarios.

2. Related work

Pedestrian detection has been a widely researched topic in computer vision, with copiousapproaches proposed to improve the accuracy and efficiency. This section provides an overview of the related work that has contributed to the advancements in pedestrian detection [1]. The HOG method, introduced by Dalal and Triggs, has been a seminal technique in pedestrian detection. HOG captures local gradient information and has proven to be an effective feature descriptor for representing pedestrians [2]. Several extensions have been proposed to enhance the HOG method, such as the inclusion of spatial pyramid representations and the combination with other feature descriptors.

Deep learning methods have made significant advancements in pedestrian detection. CNN-based approaches have demonstrated remarkable performance by leveraging the representation learning capabilities of deep neural networks [3]. For instance, the Single-shot Refinement Neural Network (RefineNet) and the Bi-box Regression Network (Bi-box) are two notable approaches that have shown promising results in handling scale variations and improving detection accuracy. The integration of contextual information has also been explored to enhance pedestrian detection algorithms. Various

DOI: 10.54254/2755-2721/32/20230208

algorithms have emerged to incorporate contextual cues, such as the Integral Channel Features (ICF) framework and the Latent HOG (LHOG) framework, to improve the robustness of pedestrian detection in handling occlusions, pose variations, and other challenging scenarios [4]. Benchmark datasets have played a critical role in evaluating and comparing pedestrian detection methods. The availability of benchmark datasets such as the PASCAL VOC dataset, the INRIA Person Dataset, the ETHZ Pedestrian Dataset, and the KITTI Pedestrian Dataset has provided valuable resources for evaluating pedestrian detection algorithms in diverse scenarios, encompassing variations in background, occlusion, and scale [5].

In summary, pedestrian detection research has seen significant advancements in feature-based methods, the availability of benchmark datasets, the adoption of deep learning techniques, and the integration of contextual information [6]. These works have collectively contributed to the development of more accurate and robust pedestrian detection systems.

3. Methodology

This section describes the methodology employed for pedestrian detection, including the theoretical principles and algorithms used in Faster R-CNN [7]. The Faster R-CNN stands as a state-of-the-art multi-stage object detector comprising two main components: a region proposal network and a region-based convolutional neural network. The architecture of Faster R-CNN is as Figure 2.

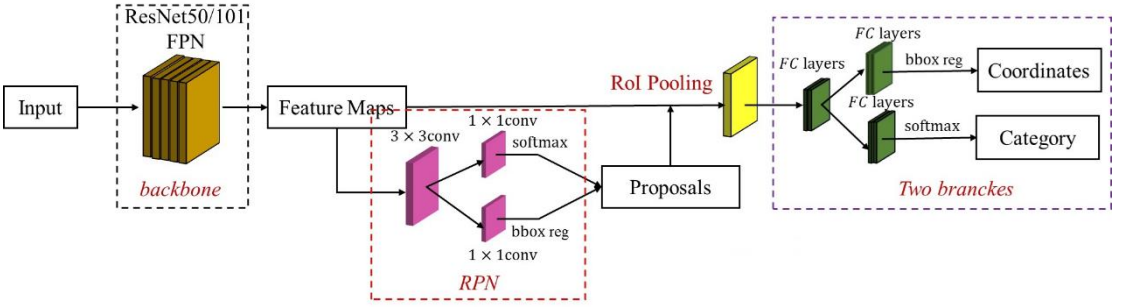


Figure 2. Faster R-CNN architecture (Photo/Picture credit: Original).

3.1. Feature extraction

The faster R-CNN uses a backbone convolutional network, usually the resnet50, to extract high-level features.

3.2. Generate proposals

After getting the feature maps, the RPN generates a set of candidate object proposals and the corresponding anchor box regression offset, which are regions likely to contain pedestrians.

3.3. Classification and bbox regression

The R-CNN component of Faster R-CNN takes the proposed regions and performs feature extraction, followed by classification and bounding box regression to accurately identify and localize pedestrians.

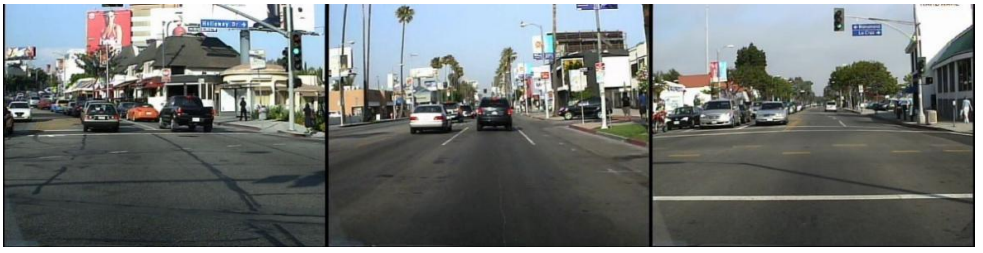
4. Experiment and analysis

4.1. Dataset

The training and testing process is based on the Caltech Pedestrian dataset, consists of 11 sets. The training sets are from set00 to set 05, and the testing set are from set06 to set10. There are more than 200,000 high-resolution images captured from vehicles driving in an urban environment in the dataset. The caltech pedestrian dataset provides the coordinates of all bounding box for pedestrians. And training images are augmented using techniques such as random flipping, scaling, and cropping to enhance the models' ability to generalize [8].

DOI: 10.54254/2755-2721/32/20230208

In the unprocessed training set, there are a lot of images that do not contain any person in it. To accelerate the training process, such images are deleted from the training set. After pre-processing, the number of training images is reduced to about 53,000. For the testing set, 1000 images are selected to get the performance of the trained model. As shown in Figure 3.



(a) (b) (c)

Figure 3. Instances of Caltech Pedestrian dataset (Photo/Picture credit: Original).

4.2. Experimental setup

4.2.1. Training section. To train the Faster R-CNN models on the Caltech Pedestrian training set, the pre-trained Resnet50-FPN is used in this experiment, to extract features of the original image. The SGD optimizer is used, with learning rate equaling 0.005, momentum equaling 0.9, weight decay equaling 0.0005. The batch size and the number of epochs are set to be 4 and 50, respectively. In each epoch, only 500 batches are used for training, to avoid over-fitting [9].

4.2.2. Testing section. In the testing section, the model is evaluated by calculating the average precision, inference time per image and resources consumption. Both training and testing section are executed on a 3070 laptop gpu, which has a video storage of 8 GB. All of the codes are implemented by pytorch framework [10].

4.3. Results and analysis

4.3.1. Training section. After 50 epochs, the total losses of each epoch are shown in the figure 4.

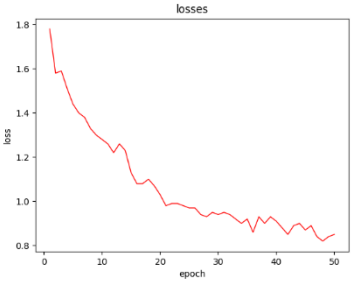


Figure 4. Total losses of each epoch (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/32/20230208

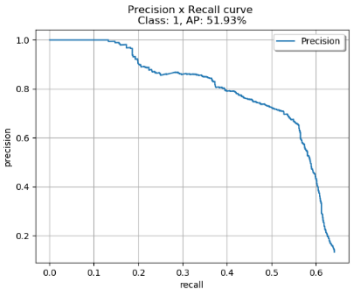


Figure 5. Precision-recall curve (epoch 45) (Photo/Picture credit: Original).

4.3.2. Testing section. In this section, 10 trained models (seperately at epoch 5,10,15, ...,50) are evaluated. Here are the average precision of models of the 10 epochs. As shown in Table 1.

Table 1. The average precision of 10 specific epochs.

Epoch	5	10	15	20	25	30	35	40	45	50
AP(%)	35.9	41.3	44.0	44.5	48.4	45.9	47.0	49.3	51.9	48.4

The model of epoch 45 can represent the best model of the whole 50 epochs, which has the AP of 51.9% in the 1000 test images. The precision-recall curve is shown Figure 5. The average inference time is 0.07s per image and the the storage occupied by the model is 158 MB, which means that the model can be used in some real-time pedetrian detection systems. Furthermore, the model could make full use of the cpu and gpu to get the better performance.

5. Discussion

Given the above results, it is noted that the model in focus achieved an Average Precision (AP) of 51.9% on the testing set. While this might appear relatively low, it is critical to take into account the specific challenges and limitations encountered during the study.

The obtained AP could potentially be attributed to the inherent complexity and diversity of pedestrians present within the dataset. Pedestrians in the dataset exhibited a wide range of poses, scales, and levels of occlusion, making accurate detection across diverse scenarios a significant challenge. Furthermore, a potential class imbalance or insufficient representation of certain pedestrian attributes might have hampered the model's capacity for generalization. Despite an AP that may appear only moderate, the model demonstrated notable inference speed, processing each image within a brisk timeframe of 0.07 seconds. This capability for real-time processing makes it a fitting candidate for use in applications requiring rapid response times, such as surveillance systems and autonomous vehicles. In terms of storage requirements, the model proved to be quite compact, occupying a mere 158MB. Such a reduced memory footprint is a beneficial trait when considering the deployment of the model on devices constrained by resources. To enhance the model's performance, there are various potential strategies to explore. For instance, the incorporation of more advanced architectural designs like attention mechanisms could facilitate the capture of more contextual information, potentially leading to increased detection accuracy. A more meticulous fine-tuning of hyperparameters, along with an exhaustive grid search, could further optimize the performance of the model.

6. Conclusion

In conclusion, the pedestrian detection model has shown a performance yielding an AP value of 51.9%. While this AP value can be viewed as moderate, it is essential to underline the model's impressive real-time inference speed and compact storage size, adding considerable value to its practical applications.

DOI: 10.54254/2755-2721/32/20230208

The resulting AP emphasizes the intricate difficulties posed by the dataset's multifaceted complexity and the diversity of pedestrian instances. Addressing these intricacies through enhanced architectural designs bears potential for augmenting the model's detection accuracy. Future investigations are encouraged to delve into more advanced techniques to boost model performance. This research opens avenues to scrutinize pedestrian detection for real-time applications. The overarching objective remains the enhancement of safety and efficiency in an array of scenarios, encompassing surveillance and autonomous systems. The insights garnered from this study establish a solid foundation for future progression in pedestrian detection, marking a valuable contribution to the burgeoning field of computer vision applications in real-world settings.

References

[1] Liu, S., Huang, D., Wang, C., & Wang, X. (2020). High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection. arXiv preprint arXiv:2005.13662.

[2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in Neural Information Processing Systems (pp. 91-99).

[3] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 886-893).

[4] Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Single-shot refinement neural network for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4203-4212).

[5] Zhu, Y., Chen, C., & Lu, H. (2019). Bi-box regression for pedestrian detection and occlusion estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3674-3683).

[6] Liu T, Stathaki T. Faster R-CNN for robust pedestrian detection using semantic segmentation network[J]. Frontiers in neurorobotics, 2018, 12: 64.

[7] Dai X, Hu J, Zhang H, et al. Multi-task faster R-CNN for nighttime pedestrian detection and distance estimation[J]. Infrared Physics & Technology, 2021, 115: 103694.

[8] Yu W, Kim S, Chen F, et al. Pedestrian Detection Based on Improved Mask R-CNN Algorithm[C]//Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020. Springer International Publishing, 2021: 1515-1522.

[9] Zhai S, Dong S, Shang D, et al. An improved faster R-CNN pedestrian detection algorithm based on feature fusion and context analysis[J]. IEEE Access, 2020, 8: 138117-138128.

[10] Zhao Z, Ma J, Ma C, et al. An improved faster R-CNN algorithm for pedestrian detection[C]//2021 11th international conference on information technology in medicine and education (ITME). IEEE, 2021: 76-80.