International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Automated Image Captioning Using Deep Learning

*Chandra .B[1] , Avinash .P[2], Sai Prasath. P[3], Jennet Shinny .D[4], Keshav Adhitya .M[5]

Submitted: 27/01/2024 **Revised:** 05/03/2024 **Accepted:** 13/03/2024

Abstract*:* Object detection, pivotal in computer vision, spans diverse applications like autonomous driving, medical imaging, etc. Deep learning, notably, enhances detection by hierarchically representing data. Two prevalent approaches are region proposal-based (e.g., R-CNN, Fast R-CNN) and unified pipeline-based (e.g., YOLOv2). The latter, exemplified by YOLOv2, emphasizes speed and simplicity. Innovations like batch normalization and anchor boxes refine accuracy. Variants like real-time YOLO adapt for specific platforms (e.g., Non-GPU computers), while methods like SSD and DSSD optimize speed and accuracy trade-offs. Recent advancements include YOLOv3's binary cross-entropy loss for improved small object detection

Keywords: *Object Detection, Deep Learning, Computer Vision, YOLOv3, Convolutional Neural Networks*

1. Introduction

The essence of identifying and categorizing items within images is fundamental across various fields, including autonomous driving, medical imaging, and surveillance. The advent of deep learning techniques has ushered in a new era in this domain, significantly enhancing the reliability, precision, and processing efficiency of these tasks.

Within the realm of object detection, two predominant methodologies reign supreme: the unified pipeline-based approach and the region proposal-based approach. In the latter, exemplified by methods like R-CNN, initial region proposals are generated within an image, followed by the extraction of features and classification using convolutional neural networks (CNNs). While these methods boast remarkable accuracy, their intricate multi-step structure contributes to complexity and time consumption.

Conversely, unified pipeline approaches prioritize simplicity and efficiency over absolute precision. These approaches, exemplified by YOLOv2, directly predict object positions and classes in a single forward pass through a CNN. YOLOv2 distinguishes itself through innovations like batch normalization and anchor boxes, enhancing overall performance. Various adaptations of YOLOv2 cater to specific needs, from real-time detection on non-GPU systems to deployment on embedded devices, showcasing its versatility and adaptability across diverse environments.

Moreover, notable advancements include Complex-YOLO, which bolsters the speed of 3D object detection, and SSD (Single Shot MultiBox Detector), which strikes a harmonious balance between speed and accuracy through the generation of multi-scale feature maps. Continual innovation in object detection methodologies, exemplified by techniques like YOLOv3 with binary cross-entropy loss, underscores a commitment to addressing challenges, particularly regarding the detection of small objects. These advancements collectively propel the field forward, enabling more effective and efficient object detection solutions tailored to specific application requirements.

2. Objective

The primary objective of this project is to implement object detection with bounding boxes using deep learning algorithms and subsequently enhance the performance analysis of the detection system. Object detection with bounding boxes is a critical task in computer vision, providing spatial context by identifying and localizing objects within frames. The inclusion of bounding boxes aids in precisely delineating the boundaries of detected objects, facilitating their recognition and interpretation by downstream applications.

Deep learning algorithms have emerged as the cornerstone of modern object detection systems, leveraging the power of artificial neural networks to learn hierarchical representations of visual data. The implementation of deep learning algorithms for object detection involves several key steps, including selecting an appropriate network architecture, fine-tuning hyperparameters, and optimizing loss functions.

Once the object detection model is trained, it can be deployed for inference to detect objects and predict their bounding boxes in real-time or batch mode. The performance of the detection system is evaluated using precision, recall, average precision (AP), and mean average precision (mAP). These quantify the accuracy and robustness of the model in detecting objects and delineating their boundaries with bounding boxes.

Enhancing the performance analysis of the object detection system involves refining existing methodologies, metrics, and evaluation techniques to provide deeper insights into the model's capabilities and limitations. Strategies for performance analysis enhancement include metric expansion, benchmarking, real-world evaluation, qualitative assessment, and generalization studies. By

Information Technology Department, Easwari Engineering College, Ramapuram, Chennai, 600089, Tamil Nadu, India.
chandra.klnce@gmail.com
avinashsarathi1712@gmail.com
saiprasath314@gmail.com
jennetshinny@gmail.com
keshavadhityam@gmail.com
** Corresponding Author Email: chandra.klnce@gmail.com*

systematically addressing each aspect of the objective, researchers and practitioners can develop more accurate, robust, and versatile object detection systems capable of meeting the demands of various real-world applications.

3. Problem Statement

The efficient detection and precise localization of objects within images or video frames are critical tasks in computer vision, with widespread applications ranging from surveillance and autonomous driving to medical imaging and industrial automation. However, achieving accurate and robust object detection with bounding boxes poses several challenges.

One significant challenge is the complexity of real-world scenes, which often contain cluttered backgrounds, occlusions, varying lighting conditions, and object scale variations. These factors can confound object detection algorithms, leading to false positives, missed detections, or inaccurate bounding box predictions. Additionally, the computational demands of deep learning-based object detection models can hinder real-time performance, particularly on resource-constrained devices or in latency-sensitive applications.

Furthermore, selecting the most appropriate deep learning architecture and optimizing its hyperparameters for a specific object detection task can be non-trivial, requiring extensive experimentation and computational resources. Moreover, accurately evaluating the performance of object detection systems poses

its own set of challenges, as existing evaluation metrics may not fully capture aspects such as bounding box quality, localization accuracy, and generalization across diverse datasets and environmental conditions.

4. Literature Survey

4.1. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model, 2020.

Advancements in object detection algorithms emphasize improving both accuracy and speed, driven by extensive research in deep learning. These algorithms have significantly enhanced object recognition across various domains like medical imaging, pedestrian detection, and autonomous vehicles. This study categorizes object detection methods into two groups: one-stage and two-stage detectors, offering a comprehensive overview. Two-stage detectors prioritize accuracy, as seen in RCNN variants, while one-stage detectors prioritize speed and efficiency, exemplified by YOLO versions. Specifically, this paper analyzes YOLOv3-Tiny, an optimized YOLO model, comparing it graphically with previous detection and recognition methods.

4.2. Automated Image Capturing System for Deep Learning based Tomato Plant Leaf Disease Detection and Recognition, 2018.

The advent of sophisticated smart farming systems, harnessing cutting-edge technologies, has revolutionized agricultural practices, particularly in the realm of tomato cultivation. Nevertheless, the delicate nature of tomato plants renders them vulnerable to an array of diseases, intricately influenced by environmental variables such as soil quality and sun exposure. In response to this pressing challenge, this study endeavors to tackle the task of disease detection within tomato crops through the application of computer vision and deep learning methodologies. Introducing an ingenious approach, the study unveils a motorized image acquisition enclosure adept at capturing comprehensive visuals of tomato plants from all angles, facilitating precise disease detection and recognition.

The system has been meticulously crafted to discern specific ailments, including Phoma Rot, Leaf Miner, and Target Spot, drawing from a dataset comprising 4,923 meticulously curated images depicting both afflicted and healthy tomato plant leaves. Employing Convolutional Neural Networks (CNNs), the system has demonstrated remarkable prowess, with an anomaly detection component achieving an 80% confidence score, and a Transfer Learning model for disease recognition boasting an impressive score of 95.75%. Moreover, the automated image capture system has been seamlessly integrated, achieving a commendable 91.67% reliability in the identification of tomato plant leaf diseases.

4.3. Design and Implementation of High Speed Background Subtraction Algorithm for Moving Object Detection, 2018

Within the expansive realm of computer vision, which encompasses a diverse array of applications ranging from surveillance and vehicle navigation to individual tracking, the task of object detection emerges as both essential and complex. Its significance reverberates throughout various sectors, where it serves as a cornerstone for maintaining public safety and combating threats such as terrorism, particularly within the intricate landscape of video surveillance. At its core, the efficacy of object detection hinges upon the ability to decipher behavioral patterns and discern moving entities within the dynamic canvas of video streams.

In this multifaceted arena, background subtraction emerges as a venerable technique, wielding the power to delineate foreground objects from their ambient backdrop with precision. In the context of this study, we present a novel and expeditious background subtraction algorithm meticulously tailored to cater to the nuances of motion-based object detection. The journey commences with the segmentation of the video into distinct streams, paving the way for subsequent processing steps. A critical component of this algorithm lies in the application of a convolution filter, strategically employed to attenuate high-frequency noise and bestow upon the imagery a semblance of fluidity and coherence.

Having smoothed the visual landscape, the algorithm proceeds to employ an adaptive background subtraction methodology, which operates dynamically to refine the delineation process. This adaptive approach imbues the system with the flexibility needed to adapt to the evolving nuances of the scene, thereby enhancing its robustness and efficacy in object detection tasks. Through the fusion of these meticulously orchestrated steps, our algorithm endeavors to push the boundaries of motion-based object detection, paving the way for enhanced surveillance capabilities and bolstered security measures.

Multi-View 3D Object Detection Network for This endeavor is dedicated to achieving precise 3D object detection within the context of autonomous driving scenarios. Enter Multi-View 3D networks (MV3D), a pioneering sensory-fusion framework that harnesses the power of both RGB images and LIDAR point cloud data to forecast oriented 3D bounding boxes. By encoding the sparse 3D point cloud through a succinct multi-view representation, MV3D orchestrates a symphony of computational prowess.

At its core, the MV3D network comprises two distinct subnetworks, each bearing its unique purpose. The first, tasked witha 3D object proposal creation, operates with commendable efficiency, crafting 3D candidate boxes through the astute utilization of aerial view representations derived from the 3D point cloud. Meanwhile, the second subnetwork orchestrates a grand fusion of multi-view features, deftly weaving together region-wise characteristics from myriad viewpoints and fostering communication across intermediate levels of diverse pathways.

Moreover, the adaptability of this approach shines through in its ability to seamlessly accommodate updates in width and height values, simplifying the process of updating cluster centers. These strides represent but a fraction of the comprehensive dataset culled from exhaustive databases, underscoring the magnitude of progress in this field.

Nevertheless, amidst these notable advancements, it remains imperative to acknowledge the inherent limitations of this approach. Chief among these concerns is the potential for escalating computing costs, which, if left unchecked, could precipitate protracted delays in item detection—a formidable challenge that warrants careful consideration.

In the proposed system, the focal point pivots to the intricate realm of object identification and localization, where a myriad of approaches vie for dominance, each wielding its unique blend of speed, accuracy, and performance. While the declaration of one algorithm's supremacy over another proves elusive, the discerning practitioner is empowered to select the approach that best aligns with the exigencies of the task at hand.

Given the expansive breadth of research within this domain, object detection applications have garnered substantial traction, yet the journey of exploration remains far from complete. Within the confines of this study, diverse algorithms for object identification and localization, varying in input image sizes, are subjected to rigorous scrutiny.

5. Proposed System

In the suggested system, object identification and localization are required. Various approaches are available, each compromising speed, accuracy, and performance. We are unable to declare one algorithm to be superior to another, though. There is always the option to choose the approach that best meets the needs. Due to the field's broad span of research, object detection applications gained a lot of traction quickly and there is still much to learn about them. In this study, different algorithms for object identification and localization with different input image sizes are compared with respect to accuracy, time, and parameter values. Our methodology increases speed without significantly compromising accuracy.

In the proposed system, the task involves identifying and pinpointing the location of objects. Numerous methods exist for this purpose, each presenting a trade-off between speed and accuracy. However, it's challenging to proclaim any single algorithm as superior to others definitively. The optimal choice depends on the specific requirements of the application. Object detection applications have gained significant traction in a short span, and there's still much ground to cover due to the expansive nature of research in this domain.

Our aim is to provide a comprehensive comparison of various algorithms for object identification and localization, considering factors such as accuracy, processing time, and parameter configurations across different input image sizes. Through our analysis, we've uncovered a novel approach utilizing a single-stage model, which significantly enhances speed without substantial compromises on accuracy.

Our comparative study reveals that YOLO v3-Tiny stands out by notably boosting the speed of object detection while maintaining satisfactory levels of accuracy. Furthermore, we propose extending the capabilities of object localization and recognition from static images to dynamic sequences, such as videos. This extension opens up exciting possibilities for real-time applications and further enhances the versatility of object detection systems.

5.1. Data Preprocessing:

Following the meticulous selection of data, a series of preprocessing procedures are meticulously executed. These encompass the transformation of video data into individual images, undertaken with utmost precision, and the subsequent reading of said images through the venerable imread() function. This pivotal phase ensures the immaculate formatting of input data, thus rendering it primed and poised for the ensuing stages of processing and analysis.

5.2. YOLOv3:

It is a veritable epitome of advancement within the annals of object detection. Harnessing the intrinsic power of probit analysis, this model endeavors to compute the targetness score for each bounding box, thereby facilitating the nuanced realm of multilabel classification. The Darknet-53 architecture, comprising a formidable ensemble of 53 convolutional layers, stands as the backbone for both feature extraction and prediction, emblematic of YOLOv3's unparalleled prowess in the domain.

5.3. Data Splitting:

The venerable module of data splitting embarks upon its solemn duty, partitioning the expanse of available dataset into two distinct cohorts: the training set and the testing set. This judicious division, an indispensable facet of model development and evaluation, serves as a crucible wherein the mettle of the trained model is tested against the unseen vistas of new data, thus illuminating the panorama of generalization prowess.

5.4. Deep Learning:

The pantheon of deep learning techniques, epitomized by the venerable artificial neural networks (ANNs), emerges as the vanguard within the system's classification endeavors. Configured with meticulous care and precision, these ANNs stand poised at the threshold of pattern recognition, their synaptic connections meticulously honed through a process of learning aimed at optimizing performance metrics.

5.5. Result Generation:

As the proverbial curtain draws near, the hallowed halls of result generation echo with the culmination of the system's noble pursuits. Here, amidst a tapestry woven from the predictions of the trained model, the final output emerges, a testament to the system's prowess in both classification and prediction realms. Performance evaluation, ensconced within the hallowed chambers of accuracy and precision, serves as the harbinger of validation, casting its discerning gaze upon the efficacy of the proposed approach.

6. Summary

The paper presents a comprehensive exploration of object detection methodologies, with a focus on two-stage and one-stage detectors. Two-stage detectors prioritize accuracy and include RCNN, Fast RCNN, and Faster RCNN, while one-stage detectors prioritize speed and efficiency and include YOLOv1, v2, v3, and

SSD. The paper specifically investigates YOLOv3-Tiny, an optimized version of the YOLO model, and compares it graphically with previous methods for object detection and recognition. The implementation methodology encompasses data selection from the MS COCO dataset, preprocessing, utilization of YOLOv3 for object detection, data splitting for training and testing, deep learning techniques for classification, and result generation. Through this methodology, the paper provides a comprehensive framework for object detection and recognition tasks, integrating various components seamlessly. Additionally, the study addresses the importance of object detection across multiple domains, highlighting its relevance in fields such as health, education, agriculture, and more. Overall, the paper contributes to the understanding of object detection methodologies and offers insights into advancements in the field, particularly focusing on the balance between accuracy and efficiency in detection algorithms.

Author contributions

Chandra B: Overall Monitoring and Supervision.
Avinash P, Sai Prasath P: Implementation and Design
Jennet Shinny D, Keshav Adithya M: Drafting and Editing

Conflicts of interest

The authors declare no conflicts of interest

References

[1].A. Krizhevsky, I. Sutskever, and G. E.Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS, 2012, doi: 10.1201/9781420010749.

[2] R. L. Galvez, A. A. Bandala, E. P. Dadios, R. R. P. Vicerra, and J. M. Z. Maningo, “Object Detection Using Convolutional Neural Networks,” IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2018- October, no. October, pp. 2023–2027, 2019, doi: 10.1109/TENCON.2018.8650517.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 580–587, 2014, doi: 10.1109/CVPR.2014.81.

[4] R. Girshick, “Fast R-CNN,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards RealTime Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[6] P. Dong and W. Wang, “Better region proposals for pedestrian detection with R-CNN,” 30th Anniv. Vis. Commun. Image Process., pp. 3–6, 2016, doi: 10.1109/VCIP.2016.7805452.

[7] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, “SSD: Single Shot MultiBox Detector,” ECCV, vol. 1, pp. 21–37, 2016, doi: 10.1007/978- 3-319-46448-0.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real- time object detection,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[9] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.

[10] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv Prepr., 2018.

[11] Ding, F. Long, H. Fan, L. Liu, and Y. Wang, “A novel YOLOv3-tiny network for unmanned airship obstacle detection,” IEEE 8th Data Driven Control Learn. Syst. Conf. DDCLS, pp. 277–281, 2019, doi: 10.1109/DDCLS.2019.8908875.

[12] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” IEEE CVPR, vol. 1, pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177.

[13] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, “Going Deeper with Convolutions,” CVPR, 2015, doi: 10.1108/978-1-78973-723-320191012.

[14] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, 2013, doi: 10.1007/s11263-013-0620- 5.

[15] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, “Object Detection with Deep Learning: A Review,” IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” ECCV, pp. 346–361, 2014, doi: 10.1023/B:KICA.0000038074.96200.69.

[17] R. Nabati and H. Qi, “RRPN : RADAR REGION PROPOSAL NETWORK FOR OBJECT DETECTION IN AUTONOMOUS VEHICLES,” IEEE Int. Conf. Image Process., pp. 3093–3097, 2019.

[18] L. Jiao et al., “A Survey of Deep Learning-Based Object Detection,” IEEE Access, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/access.2019.2939201.

[19] D. Wang, C. Li, S. Wen, X. Chang, S. Nepal, and Y. Xiang, “Daedalus: Breaking Non-Maximum Suppression in Object Detection via Adversarial Examples,” arXiv Prepr., 2019.

[20] C. Ning, H. Zhou, Y. Song, and J. Tang, “Inception Single Shot MultiBox Detector for object detection,” IEEE Int. Conf. Multimed. Expo Work. ICMEW, no. July, pp. 549–554, 2017, doi: 10.1109/ICMEW.2017.8026312.

[21] Z. Chen, R. Khemmar, B. Decoux, A. Atahouet, and J. Y. Ertaud, “Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility,” 8th Int. Conf. Emerg. Secur. Technol. EST, pp. 1–6, 2019, doi: 10.1109/EST.2019.8806222.

[22] D. Xiao, F. Shan, Z. Li, B. T. Le, X. Liu, and X. Li, “A Target Detection Model Based on Improved Tiny-Yolov3 Under the Environment of Mining Truck,” IEEE Access, vol. 7, pp. 123757–123764, 2019, doi: 10.1109/access.2019.2928603.

[23] Q. C. Mao, H. M. Sun, Y. B. Liu, and R. S. Jia, “Mini-YOLOv3: RealTime Object Detector for Embedded Applications,” IEEE Access, vol. 7, pp. 133529–133538, 2019, doi: 10.1109/ACCESS.2019.2941547.

[24] W. Fang, L. Wang, and P. Ren, “Tinier-YOLO: A Real-time Object Detection Method for Constrained Environments,” IEEE Access, vol. 8, pp. 1935–1944, 2019, doi10.1109/ACCESS.2019.2961959.

[25] R. Huang, J. Pedoeem, and C. Chen, “YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers,” IEEE Int. Conf. Big Data, Big Data, pp. 2503–2510, 2019, doi: 10.1109/BigData.2018.8621865.