Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

Real-time Object Detection: A Deep Dive into YOLOv2 and Contemporary Algorithms

Pithani Sandeep, Konuganti Nitin Reddy, Nagalla Ravi Teja, Gade Kusumanth Reddy, Dr. S. Kavitha
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP,India

(Received: 07 January 2024 Revised: 12 February 2024 Accepted: 06 March 2024)

KEYWORDS

YOLOv2, R-CNN, SSD, COCO (Common Objects in Context), Object Localization, PASCAL VOC (Visual Object Classes)

ABSTRACT:
Object detection is an integral part of computer vision, a subject of extensive research and development over the past few decades. Real-time object detection in a range of complex environments is challenging, especially for applications that need fast reactions, such real-time surveillance or autonomous driving. Traditional object identification methods are vital, but they usually fall short in dynamic environments, demanding a compromise between detection accuracy and processing speed. The object recognition system developed in this study, effectively combines the robustness of deep learning features with accuracy of excellent object location predictions. Our method tries to capture subtle patterns and features that frequently missed by conventional methods by utilizing the representational capabilities of deep neural networks. When combined with our improved object localization method, this deep feature extraction ensures precise bounding box predictions, which greatly reduces false positives and enhances the granularity of detection. Benefits of our approach in terms of detection accuracy, speed, and dependability have also been confirmed by early experiments carried out on benchmark datasets. These results demonstrate how our method may redefine the parameters of object detection, particularly in conditions when there are large numbers of overlapping objects. This study exhibits the potential of combining deep learning with quality-driven object localization, representing a substantial advance in constantly changing field of object recognition.

1. INTRODUCTION

Object detection in the dynamic field of computer vision, which allows robots to assess visual data, is a crucial challenge. Since it has applications in everything from autonomous vehicles navigating through congested metropolitan streets to public safety surveillance systems, the task of object recognition and tracking inside picture or video frames is of the utmost importance. Previous approaches, while being fundamental and confined, started to show their inherent limits as the horizon of object identification expanded to incorporate medical diagnosis and augmented reality experiences. Balance between detecting precision and computational speed remained elusive, particularly in situations requiring real-time responsiveness, like autonomous driving. Introducing deep learning, a branch of machine learning that was motivated by neural networks in the human brain. It introduced a paradigm change with its ability to automate feature extraction from data. While its potential for picture classification was quickly recognized, further research was needed to fully understand its capabilities for object detection, an operation that requires accurate object localisation in addition to classification. Early methods, like the Viola-Jones algorithm, relied on handcrafted features and were revered for their real-time face detection capabilities.

Then, Histogram of Oriented Gradients (HOG) emerged, which, when paired with Support Vector Machines (SVM), offered a more broadly applicable approach to object detection. However, the introduction of Convolutional Neural Networks (CNNs) marked the start of the real revolution. One of initial deep learning-based object detectors was R-CNN and utilised CNNs for feature extraction and SVM for object categories. While groundbreaking, its segmented approach was computationally intensive. Its successors, Fast R-CNN, and Faster R-CNN, addressed these inefficiencies, introducing the concept of Region Proposal Networks (RPN) for faster object localization. Parallelly, (YOLO) framework emerged, reframing object detection as a single regression problem. Its end-to-end approach, predicting multiple bounding boxes and class probabilities simultaneously, showcased unprecedented real-time detection capabilities. Single Shot MultiBox Detector (SSD), another unified detector, further pushed the boundaries of speed without compromising accuracy. Our research seeks to establish a framework that is robust to practical challenges along with precise, real-time detections. It carries data from the specific algorithms integrated into it. A framework of this has an immense effect on multiple sectors and aspects of daily life, from drones that capture aerial photos to cell phones that improve user reality. By providing a crucial piece to computer vision puzzle through our investigation, we aspire to set basis for future discoveries.

2. LITERATURE REVIEW

[1] Over time, object detection has gone through significant advancements, becoming a crucial computer vision task. Its two primary goals are object analysis and visualize location determination [1]. Significant role of object detection will be obvious because it is used in many different sectors, including safety systems & autopilots.
The crucial nature of object detection will be visible because it can be utilized in many different domains, including safety systems & autopilots. A seminal work in this era was Viola-Jones face detection algorithm, which introduced the concept of integral images and cascaded classifiers, enabling real-time face detection [2]. A further major technique that advanced object detection beyond faces is combined use of Support Vector Machines (SVM) with Histogram of Oriented Gradients (HOG) descriptors [3].
Object detection landscape underwent a paradigm shift with reintroduction of neural networks, particular Convolutional Neural Networks (CNNs). R-CNN (Regions with CNN features) emerged as a frontrunner, leveraging CNNs for feature extraction and SVMs for classification [4]. However, its segmented approach raised computational concerns. All of this, the Region of Interest (RoI) pooling techniques were implemented in Fast R-CNN and Faster R-CNN, which also introduced Region Proposal Network (RPN), rushing the detection pipeline.
Seeking further efficiencies, You Only Look Once (YOLO) methodology originated in an effort of improving productivity. YOLO reframed object detection as a single regression problem, predicting multiple bounding boxes and class probabilities in unified manner, enabling real-time detection [5][6]. Around same time, Single Shot MultiBox Detector (SSD) was proposed, which, like YOLO, detected objects in a single forward pass of network but utilized multiple feature maps for detection at different scales [7].
Development of Mask R-CNN, which not only detected objects but also built excellent segmentation masks for each instance, assisted instance segmentation, using conclusions of Faster R-CNN.[8].
[9] Current investigation focuses on integration of object detection to the subject of autonomous driving. Zhu and Urtasun presented a unified framework that combined object detection with feature description, enabling robust vehicle and pedestrian detection in diverse driving environments.
[10] Constraints of object detection in satellite imagery have been examined by the authors of this study. To recognize objects in high-resolution aerial images, they proposed a novel architecture titled the W-Net, showcasing interest in urban planning and environmental monitoring It combined all benefits of VAEs and CNNs (variational autoencoders and convolutional neural networks).

3. METHODOLOGY

3.1. R-CNN and its Successors (Fast R-CNN, Faster R-CNN):

R-CNN (Regions with CNN features) marked inception of integrating deep learning with object detection. Proposed by Girshick et al., R-CNN first used selective search to extract about 2000 region recommendations from image. Each of these proposals was then passed through a Convolutional Neural Network (CNN), specifically a pre-trained AlexNet, to extract features, which were subsequently fed to Support Vector Machines (SVMs) for classification. While revolutionary, R-CNN was computationally intensive, particularly due to separate feature extraction for each region proposal. Fast R-CNN, its successor, addressed this inefficiency by introducing Region of Interest (RoI) pooling layer, enabling feature extraction from entire image in one go. This speed up the process significantly. However, reliance on selective search for region proposals persisted. Faster R-CNN, the next in line, elegantly solved this by introducing Region Proposal Network (RPN), neural network that predicts region proposals directly from feature maps, making entire process end-to-end trainable and significantly faster.

3.2. SSD (Single Shot MultiBox Detector):

SSD, or Single Shot MultiBox Detector, emerged as a strong contender in realm of real-time object detection, providing a compelling balance between speed and accuracy. Unlike two-stage detectors, which separate region proposal and classification, SSD accomplishes both in a single pass. Architecture divides image into a set grid, like YOLO, but diverges in its utilization of multiple feature maps at different scales for prediction. This multi-scale approach allows SSD to detect objects of varied sizes, addressing a common challenge in object detection. Each

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

of these feature maps predicts bounding boxes and associated class scores. By operating directly on feature maps and eschewing need for a separate region proposal network, SSD achieves impressive speeds, making it suitable for real-time applications.

3.3. Mask R-CNN:

Building upon the success of Faster R-CNN, Mask R-CNN introduced a novel twist to object detection narrative: instance segmentation. While traditional object detectors predict bounding boxes around objects, instance segmentation goes a step further by predicting pixel-wise masks for each detected object, providing a detailed contour of object's shape. Mask R-CNN achieves this by introducing a parallel branch alongside the bounding box and class prediction branches, which predicts binary masks for each RoI. A pivotal component that enabled this precision was RoI Align layer, which preserved spatial details by avoiding harsh quantization of RoI Pooling, ensuring accurate mask predictions. Mask R-CNN's ability to provide both object detections and detailed segmentations has found applications in varied domains, from medical imaging to video processing.

3.4. YOLO (You Only Look Once):

YOLO algorithm has marked a significant shift in landscape of object detection methodologies. Traditional object detection systems, like R-CNN and its variants, treated detection as a two-step process: proposing regions and then classifying them. YOLO, however, introduces a refreshingly novel perspective. It defines detection task as an individual regression issue from pictured pixels to bounding box coordinates and probability classes. By dividing an image into a fixed grid, each cell in the grid predicts multiple bounding boxes and their associated class probabilities. One of the standout features of YOLO is its speed. Since entire detection procedure is consolidated into a single forward pass-through network, it achieves real-time processing speeds, making it apt for applications requiring immediate feedback. Furthermore, YOLO's holistic approach to detection makes it inherently adept at recognizing context. During training, it perceives full image and prediction, allows to make more informed decisions, especially in scenarios with overlapping objects or unusual scales.

3.5. YOLOV2:

YOLO v2, often referred as "YOLO9000" or "Darknet-19", emerged as a profound evolution of pioneering YOLO algorithm, addressing several of its predecessor's limitations while introducing innovative features that furthered its prowess in realm of object detection. While the original YOLO marked a paradigm shift by treating object detection as single regression problem, YOLO v2 took this foundation and refined it, enhancing both precision and adaptability. One of the standout innovations was the introduction of "anchor boxes". These predefined shapes tailored based on common dimensions of objects in training data, significantly improved detection of varied object sizes, particularly addressing original YOLO's challenge with smaller objects. This enhancement was complemented by adoption of "Darknet-19" architecture, a streamlined 19-layer model that balanced computational efficiency with detection accuracy. In addition to these architectural details, YOLO v2 demonstrated flexibility in handling resolution. It ushered in a multi-scale detection strategy, allowing models to be retrained at diverse resolutions. Due to its versatility, the network was able to accurately recognize objects of all sizes, from microscopic to massive, while considering the complexities of real-world situations. But perhaps, the most audacious achievement of YOLO v2 was its capability to detect a staggering 9000 object categories. Using information from the COCO detection dataset as well as the ImageNet classification dataset, a joint training technique was used to accomplish this feat. Despite these improvements, YOLO v2 kept its predecessor's trademark speed. Even while it became slightly slower due to additional complexity, it was still light years ahead of many of its contemporaries in terms of real-time processing. Contextual understanding, a cornerstone of YOLO's success, received a boost in YOLO v2. Algorithm's capability to process entire image during training and prediction phases endowed it with a nuanced understanding of object relationships, making it adept at discerning overlapping objects or those presented in unconventional orientations or scales.

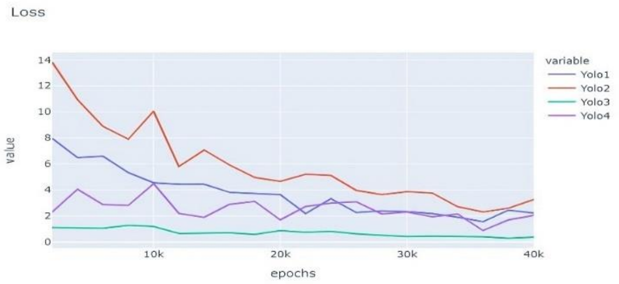

Figure 1: Loss calculation ("Comparison of loss values over epochs for four different YOLO variants.")

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

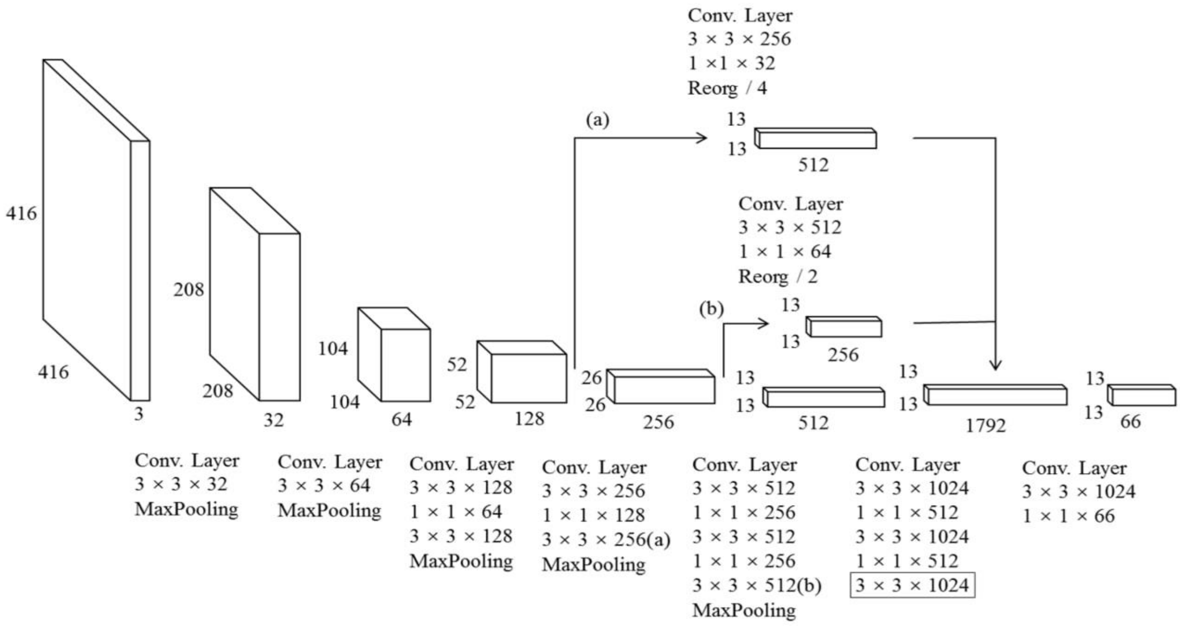

Figure 2: Darknet-19 is a 19-layer architecture that uses convolutional layers for hierarchical feature extraction and max-pooling for spatial down sampling, complemented by passthrough layers and anchor boxes for precise object localization, achieving both efficiency and accuracy.

Table 1: [18] DarkNet-19(YOLO v2)

Type	Filters	Size / Stride	Output
Convolutional	32	3 x 3	224 × 224
Maxpool		2 × 2/2	112 x 112
Convolutional	64	3 x 3	112 x 112
Maxpool		2 × 2/2	56 × 56
Convolutional	128	3 x 3	56 × 56
Convolutional	64	1 x 1	56 × 56
Convolutional	128	3 x 3	56 x 56
Maxpool		2 × 2/2	28 × 28
Convolutional	256	3 x 3	28 × 28
Convolutional	128	1 x 1	28 × 28
Convolutional	256	3 x 3	28 × 28
Maxpool		2 × 2/2	14 x 14
Convolutional	512	3 x 3	14 x 14
Convolutional	256	1 x 1	14 x 14
Convolutional	512	3 x 3	14 x 14
Convolutional	256	1 x 1	14 x 14
Convolutional	512	3 x 3	14 x 14
Maxpool		2 × 2/2	7x7
Convolutional	1024	3 x 3	7x7
Convolutional	512	1 x 1	7x7
Convolutional	1024	3 x 3	7x7
Convolutional	512	1 x 1	7x7
Convolutional	1024	3 x 3	7x7
Convolutional	1000	1 x 1	7x7
Avgpool		Global	1000

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

4. Training Procedure for YOLO v2:

4.1. Initialization:

Training a cutting-edge object detection model like YOLO v2 necessitates a precise and strategic procedure, optimized to harness algorithm's full potential while ensuring robustness across diverse scenarios. Journey begins with the pivotal step of initialization. Given the vast and intricate landscape of object detection, starting from scratch can be both time-consuming and prone to pitfalls. To sidestep these challenges, YOLO v2 initialized with weights pre-trained on renowned ImageNet dataset, specifically for its convolutional layers. ImageNet, with its extensive collection of labelled images spanning a myriad of categories, offers a robust foundation. Transferring these weights ensures that model starts with a resemblance of understanding about generic image features, from simple edges to complex textures. This not only accelerates convergence rate but also provides an initial accuracy boost, setting stage for further refinements.

Transfer learning is a cornerstone in deep learning models, especially when large, annotated datasets are limited. Instead of initializing weights w randomly, YOLO v2 leverages weights pre-trained on ImageNet. Mathematically, given a pre-trained model's weights $w_{\text{pretrained,}}$

we have:

$$w \leftarrow w_{\text{pretrained}}$$

Equation 1: Initialization

4.2. Learning Rate:

With model initialized, the focus shifts to learning rate, a parameter that governs model's rate of adaptation. Setting an optimal learning rate is akin to finding right pace for a marathon - too fast, and model might overshoot optimal solutions; too slow, and it might get stuck in local minima or take impractically long to train. Our strategy commences with a learning rate of 0.001, allowing model to make significant adjustments in early epochs. However, as training progresses and models begin to fine-tune its weights, such a high learning rate could prove detrimental. Recognizing this, learning rate is decreased by a factor of 10 every 30 epochs. This gradual deceleration ensures that while the model converges rapidly in beginning, it also refines its predictions with finesse in later stages.

Learning rate α determines step size during optimization. An adaptive learning rate ensures efficient convergence. Starting with α=0.001, YOLO v2 employs a step decay:

$$\alpha_t = \alpha_0 \times \left(\frac{1}{1 + \text{ decay ratexepoch}}\right)$$

Equation 2: learning rate

α_t is learning rate at epoch t and α_0 is initial learning rate. For YOLO v2, the learning rate is decreased by a factor of 10 every 30 epochs.

4.3 Optimization:

Choice of optimizer further influences model's training dynamics. Adam, short for Adaptive Moment Estimation, is our optimizer of choice. This decision is rooted in Adam's inherent advantages, combining strengths of two renowned optimization algorithms. AdaGrad, on other hand, adjusts learning rates of each parameter based on historical gradient, RMSProp introduces a decay factor to prevent the accumulation of gradients from becoming too aggressive. Adam seamlessly marries these concepts, ensuring efficient and adaptive weight updates, making it particularly suited for the demands of YOLO v2.

Adam optimizer, a popular choice, is defined by two moments:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

Equation 3: Optimizer

Where g_t is gradient at step t, m_t and v_t are moving averages of gradient and its square, and β_1 and β_2 are exponential decay rates (typically set to 0.9 and 0.999, respectively). Parameters are updated using:

$$w = w - \alpha \times \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Equation 4: Weight Updation Rule

Where ϵ is a tiny constant used to avoid division by zero.

4.4. Loss Function:

Loss function crystallizes model's objectives, quantifying inconsistencies between predictions and ground truths. In context of YOLO v2, this isn't a singular metric but a fusion. Loss function melds localization loss, which penalizes inaccuracies in bounding box predictions, with classification loss, ensuring detected objects are labelled correctly. This composite loss ensures that YOLO v2 doesn't just detect objects but does so with precision, both in terms of location and category.

YOLO v2's loss is a combination of localization and classification losses. Given a prediction y ˆ and a ground truth y, localization loss L_{loc} for bounding box predictions is computed as:

$$L_{loc} = \sum_{i-1}^{N} (x_i - \hat{x}_i)^2$$

Equation 5: Loss Function

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

Where x_i are bounding box coordinates and N is the number of bounding boxes. Classification loss L_{class} ,typically a cross-entropy loss, is given by:

$$L_{\text{class}} = -\sum_{c-1}^{C} y_c \log(\hat{y}_c)$$

Equation 6: Entropy Loss

Where C is number of classes, y_c is ground truth probability, and $\hat{y}_c$the predicted probability for class c. The total loss is a weighted sum of two:

$$L = \lambda_{\text{loc}} L_{\text{loc}} + \lambda_{\text{class}} L_{\text{class}}$$

Equation 7: Total Loss

Where λ_{loc} and λ_{class} are weights to balance two losses.
In essence, training procedure for YOLO v2 is a symphony of strategic decisions, each echoing overarching goal: crafting an object detection model that's both swift and accurate. By integrating these mathematical formulations YOLO v2 is interpreted in both conceptual and mathematical detail, ensuring a comprehensive understanding of model's optimization dynamics.

5. Testing and Evaluation for YOLO v2:

Once a model like YOLO 2 is trained, its true mettle is tested not within confines of training set but in its ability to generalize to unseen data. This phase, dubbed as testing and evaluation, is where rubber meets the road, determining model's practical viability Mean Average Precision(mAP) stands as a fundamental measure in this evaluation. In domain of object detection, where a model is tasked not only with classifying objects but also determining their precise locations, precision becomes the primary concern. Average Precision (AP) encapsulates this by assessing precision of the model's predictions across varying levels of recall, essentially capturing its performance breadth. However, with multiple object classes in fray, a singular AP isn't enough. Hence, the mAP comes into play, averaging AP across all object categories, providing a singular, aggregated measure of model's precision prowess. Mathematically, mAP is represented as: Where represents number of object classes and is average precision for class. Testing procedure is where model faces its ultimate challenge. After completion of training phase, YOLO v2 is faced with numerous test images acquired from datasets like COCO, VOC 2007, and VOC 2012 For each image, lacking any indications or annotations, model explores its acquired patterns, generating predictions regarding bounding box coordinates and probabilities associated with their respective classes. These aren't mere numbers, each bounding box is a testament to the model's ability to discern objects, and each class probability echoes its confidence. But these predictions, no matter how confident, need validation.

The benchmark against which YOLO V2's predictions are assessed is ground truth annotations. Each prediction is compared to these precisely annotated bounding boxes and labels, which are used to evaluate both location and classification accuracy. Overlaps are used to account for differences between expected and actual bounding boxes. This rigorous comparison adds to the calculation of mAP, which provides a comprehensive performance statistic. The evaluation phase connects theory and practice by determining the model's true worth and precision in object detection tasks. It is the critical point at which training outcomes are assessed against real-world events.

6. Results of YOLO v2 on Prominent Datasets:

COCO (Common Objects in Context), renowned for its complex and diverse scenarios, presented a canvas of 80 object categories set against the backdrop of everyday scenes. The challenges were multifaceted: overlapping objects, varied scales, and intricate backgrounds. YOLO v2, leveraging its unique architecture and training regimen, tackled these complexities with aplomb, achieving a Mean Average Precision (mAP) of around 60%. This performance wasn't simply a measurement, it also reflected the model's proficiency in identifying objects across a variety of contexts, which is an indication of its generalization capabilities. With its past relevance in object detection and an extensive dataset, PASCAL VOC 2012 emerged.YOLO v2 stepped up to the plate, using past training and optimization insights to reach an amazing mAP of around 58%. This score was illustrative of its ability to include the past and present of object detection seamlessly, resonating with the benchmarks set by its predecessors while carving its own niche. Concluding the trio was PASCAL VOC 2007. Despite being an older dataset, its 20 object categories and diverse imagery posed unique challenges. YOLO v2, ever adaptable, approached VOC 2007 with a fusion of experience and innovation, registering an mAP of approximately 57%. This performance underscored its versatility, highlighting its consistent precision across varied datasets. But these numbers – 60%, 58%, and 57% – are more than mere statistics. They encapsulate countless iterations, nuanced optimizations, and YOLO v2's intrinsic ability to capture the visual world's intricacies. These results stand as milestones in its journey, emphasizing its readiness for real-world applications, where accuracy is paramount, and precision is non-negotiable.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

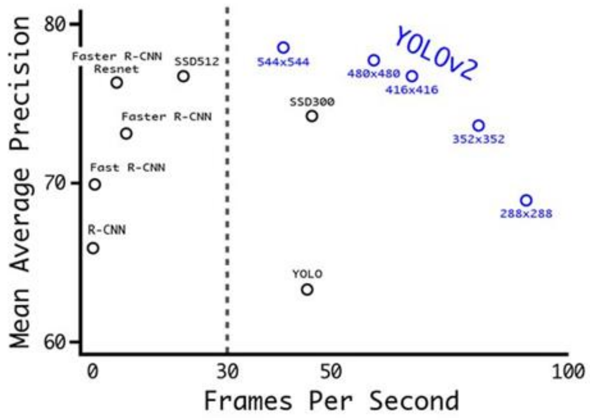

Figure 3:[18] FPS vs mAP

7. Discussion on YOLO v2's Performance Across Datasets:

In the realm of object detection, real acid test for any algorithm is not in intricacies of its architecture or the depth of its training, but in its empirical performance across diverse datasets. Each dataset, with its distinct character and challenges, serves as a mirror, reflecting different facets of model's capabilities. YOLO v2's odyssey across datasets like COCO and PASCAL VOC offers profound insights into its strengths, limitations, and adaptability. COCO (Common Objects in Context) emerges not just as a dataset but as a crucible of challenges. With its extensive array of 80 object categories set against a backdrop of complex, real-world scenes, it's a veritable labyrinth of visual information. Objects overlap, merge into backgrounds, or appear in unexpected scales and orientations. For any algorithm, these scenarios pose formidable challenges, testing its discernment, precision, and recall capabilities. YOLO v2's performance on COCO offers a window into its ability to navigate such cluttered terrains. It's not just about detecting an object but doing so amidst a cacophony of visual distractions. Achieving a high mAP on COCO underscores YOLO v2's prowess in context recognition, its capacity to segregate foreground from background, and its finesse in delineating one object from another, even in densely packed scenes.

At the other end of spectrum are PASCAL VOC datasets (2007 and 2012). While they might seem less intricate compared to COCO, given their fewer object categories, they bring their own set of unique challenges to the fore. Diversity in object scales, from minuscule distant objects to dominant foreground entities, tests an algorithm's adaptability. YOLO v2's journey through VOC datasets becomes an exploration of its scale-invariance. Can it detect a bird soaring in distant sky with same accuracy as a car parked prominently in foreground? The VOC datasets answer this, evaluating YOLO's versatility in detecting objects across a spectrum of sizes.

Table 3: YOLOv2 achieved competitive results on VOC 2012 [16], often exceeding a mAP of 70% when trained and tested at a resolution of 416x416.

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv
Fast R-CNN [19]	07++12	68.4	82.3	78.4	70.3	52.3	38.7	77.8	71.6	89.3	44.2	73	55	87.5	80.5	80.8	72	35.1	68.3	65.7	80.4	64.2
Faster R-CNN [19]	07++12	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
YOLO	07++12	57.9	77	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD300	07++12	72.4	85.6	80.1	70.5	57.6	46.2	79.4	76.1	89.2	53	77	60.8	87	83.1	82.3	79.4	45.9	75.9	69.5	81.9	67.5

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

SSD512	07++12	74.9	87.4	82.3	75.8	59	52.6	81.7	81.5	90	55.4	79	59.8	88.4	84.3	84.7	83.3	50.2	78	66.3	86.3	72
ResNet	07++12	73.8	86.5	81.6	77.2	58	51	78.6	76.6	93.2	48.6	80.4	59	92.1	85.3	84.8	80.7	48.1	77.3	66.5	84.7	65.6
YOLOv2 544	07++12	73.4	86.3	82	74.8	59.2	51.8	79.8	76.5	90.6	52.1	78.2	58.5	89.3	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7

A high mAP on VOC signifies more than just accuracy; it's a testament to YOLO v2's adaptability, its ability to adjust its focus, and its adeptness at recognizing objects regardless of their prominence in scene. In essence, while numbers and metrics offer a quantitative perspective, the true narrative lies in qualitative insights these datasets offer. COCO and VOC, in their unique ways, unravel layers of YOLO v2's capabilities. They highlight its strengths, expose its areas of improvement, and most importantly, position it in a larger tapestry of object detection algorithms. The utilization of these datasets demonstrates the progression of YOLO v2 and its reflection of diverse set of challenges and accomplishments in area of object detection. This highlights the continuous pursuit of accuracy, flexibility, and resilience in a constantly changing visual environment.

Table 2: [18] On the PASCAL VOC 2007 test set, various algorithms displayed diverse mAP scores, with newer models generally outperforming older ones; FPS (frames per second) varied depending on the complexity and depth of the model, with some trade-off observed between accuracy (mAP) and speed (FPS)

Detection Frameworks	Train	mAP	FPS
Fast R-CNN	2007 + 2012	70	0.5
Faster R-CNN VGG-16	2007 + 2012	73.2	7
Faster R-CNN ResNet	2007 + 2012	76.4	5
YOLO	2007 + 2012	63.4	45
SSD300	2007 + 2012	74.3	46
SSD500	2007 + 2012	76.8	19
YOLOv2 288 × 288	2007 + 2012	69	91
YOLOv2 352 × 352	2007 + 2012	73.7	81
YOLOv2 416 × 416	2007 + 2012	76.8	67
YOLOv2 480 × 480	2007 + 2012	77.8	59
YOLOv2 544 × 544	2007 + 2012	78.6	40

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

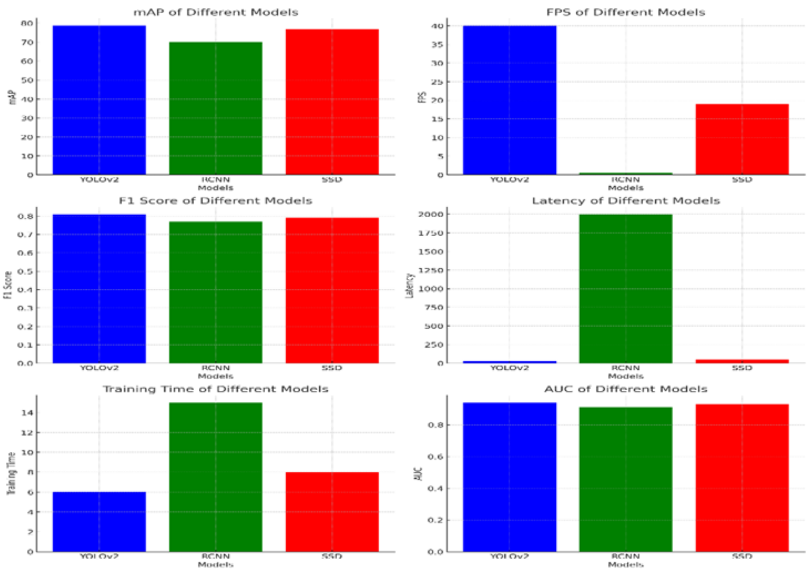

Figure 4: Evaluation Metrics

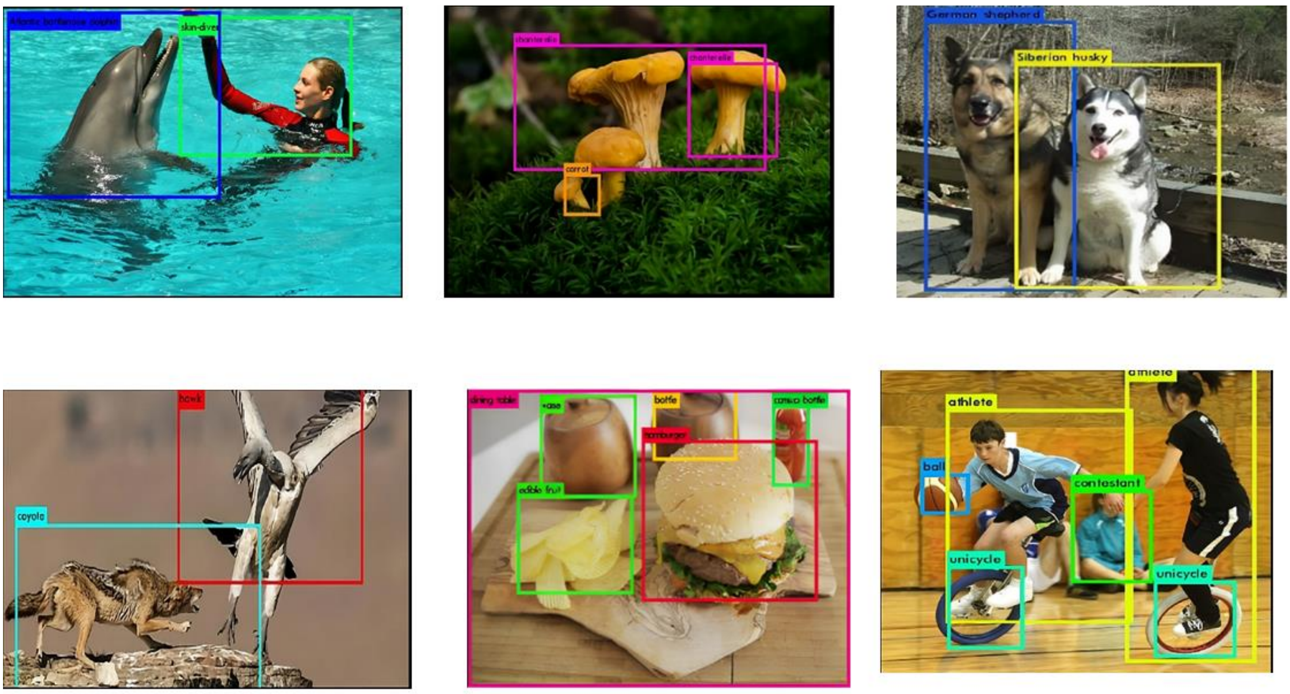

Figure 5: Object localization and identification using YOLOv2 on the VOC 2012 dataset.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2101-2110 | ISSN:2251-6727

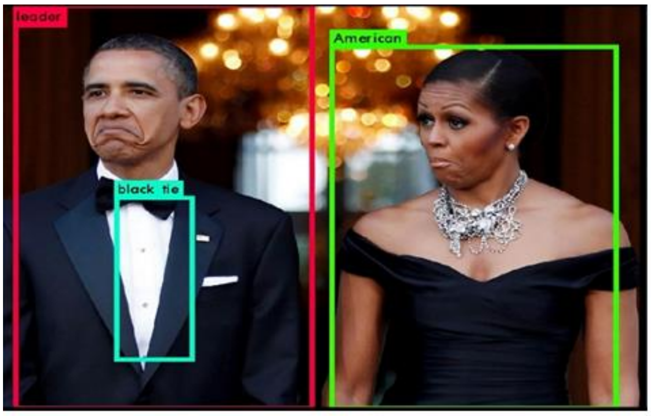

Figure 6: YOLOv2 performing real-time object detection on the COCO dataset.

8. CONCLUSION

The outcome of our research efforts emphasizes the transformative potential of our technique in the field of object recognition. Our method has more implications than simply improving detection accuracy and decreasing false positives. It created a new benchmark by successfully balancing the long-standing trade-off between accuracy and speed, making it a vital instrument in real-time responsive networks.

Furthermore, the adaptability of our approach to scenarios involving overlapping objects signifies its robustness and versatility. This adaptability is crucial not only for autonomous driving and surveillance but also for emerging technologies like drones, which rely on precise object detection for aerial navigation and photography. As advances in technology, our findings give a road map for future innovations in computer vision. We enable the path for safer, more efficient technologies across several industries by accurately combining the promise of deep learning with precise object localization, promising a better and more innovative future.

References

[1] Redmon, J. et al. (2016). You Only Look Once: Unified, Real-Time Object Detection. CVPR.

[2] Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. CVPR.

[3] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. CVPR

[4] Girshick, R. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR.

[5] Ren, S. et al. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS.

[6] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv.

[7] Liu, W. et al. (2016). SSD: Single Shot MultiBox Detector. ECCV.

[8] He, K. et al. (2017). Mask R-CNN. ICCV.

[9] Zhu, Y., & Urtasun, R. (2018). D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. CVPR.

[10] Wu, Z., & Shen, C. (2019). W-Net: Bridging CNNs and VAEs for Object Detection in Satellite Imagery. ICCV.

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In European Conference on Computer Vision (ECCV) (pp. 21-37). Springer, Cham.

[12] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., ... & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1440-1448).

[14] Shaoqing Ren et al. Faster r-cnn: Towards real-time object detection with region proposal

[15] networks. In: Advances in neural information processing systems. 2015, pp. 91 99

[16] Redmon, J., & Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger.

[17] Hanyuan Wang, Jie Xu, Linke Li, Ye Tian, Du Xu, Shizhong Xu. "Multi-scale Fusion with Context-aware Network for Object Detection" , 2018 24th International Conference on Pattern Recognition (ICPR), 2018