Detailed analysis on the existing methods of OBJECT DECTECTION

Vasa Koushik, Repaka Lokesh, Srikanthi Chadalavada

SRM University, Andhra Pradesh

Abstract:

One of the most fundamental and difficult tasks in computer vision is object detection, which aims to locate object instances in natural images from a huge number of established categories. Deep learning approaches have developed as a potent tool for learning feature representations directly from data, resulting in significant advances in the field of generic object detection. Given this period of rapid evolution, the purpose of this work is to present a complete overview of current deep learning-related advances in this subject. This survey includes nearly 300 research contributions that span a wide range of general object detection topics, including detection frameworks, object feature representation, object proposal generation, context modelling, training methodologies, and evaluation metrics.

A research analysis by spotting the promising directions for upcoming research.

Introduction:

Object detection is a computer vision approach for detecting things in photos and videos. To obtain relevant results, object detection algorithms often use machine learning or deep learning. When we look at photographs or videos, we may quickly distinguish and find objects of interest. The purpose of object detection is to use a computer to imitate this intelligence.

If we take **Image categorization**, for example, is straightforward, but the distinctions between object localisation and object detection can be perplexing, especially when all three tasks are referred to as object recognition. Object localization entails creating a bounding box around one or more objects in an image, whereas image classification entails providing a class name to an image. Object detection is more difficult, as it combines the two tasks by drawing a bounding box around each object of interest in the image and assigning a class name to it. Object recognition is the collective term for all of these issues.

Generic object detection, also called generic object categorydetection, object class detection, or object category detection (Zhang et al. 2013). It determines whether or if there are instances of objects from predefined categories in an image (often many categories, such as 200 in the ILSVRC object detection challenge) and, if so, return the spatial location and extent of each instance.

Detecting a broad range of natural categories is prioritised over particular item category detection, which may only detect a restricted predetermined category of interest (e.g., faces, people, or autos). Despite the fact that the visual world in which we live contains thousands of objects, the research community is currently focused on the localization of highly structured objects (e.g., cars, faces, bicycles, and aeroplanes) as well as articulated objects (e.g., humans, cows, and horses) rather than unstructured scenes.

Generic Object Detection is closely related to semantic image segmentation. There are numerous issues that are strongly related to generic object detection. The purpose of object classification, also known as object categorization, is to determine the existence of items from a given set of object classes in an image; that is, assigning one or more object class labels to a given image and determining presence without the requirement for location. Detection is more difficult than classification due to the extra requirement of locating instances in a picture. The object recognition problem encompasses the more broad challenge of identifying/localizing all of the items in an image, as well as the object detection and classification difficulties. (Everingham et al. 2010; Russakovsky et al. 2015; Opeltetal. 2006; Andreopoulos and Tsotsos 2013).

This research work plays a very important role in describing the main topic of the research domain. The research domain in Deep Learning, which is a part of Machine Learning.

Coming to the research point of view, this analysis helps in creating a clear and precise understanding of the concepts involved in the techniques used for Object Detection for future research and also provides a theoretical edge to Corporate World which can reduce consuming time as this research analysis includes detailed explanation.

As, mentioned above alot of researchers, scholars and students will have a standard learning and fills the gap of understanding this concept.

This is a research work which give an insight to the existing object Detection and provide an analysis on all the techniques inclused and their features.

Related works: presents review of the previous work on this topic:

There are numerous number of important reviews based on object diction being published. But the most recognised are based on problem of specific object detection like pedestrian detection, face detection, vehicle detection and text detection. And comparatively there are very few reviews focuses specifically on generic object detcetion. Apart from the work of Zhang et al.(2013) carried out a survey on object class detection.

Deep learning enables computational models to learn fantastically complex, subtle, and abstract representations, enabling significant progress in a variety of fields including visual recognition, object detection, speech recognition, natural language processing, medical image analysis, drug discovery, and genomics. DCNNs (LeCun et al. 1998, 2015; Krizhevsky et al. 2012a) are one of the different forms of deep neural networks that have made advancements in image, video, voice, and audio processing.

In contrast, despite the fact that various deep learning-based algorithms for object detection have been presented, we are unaware of any comprehensive current algorithms . For further advancement in object detection, a detailed review and explanation of prior work is required, especially for researchers new to the subject. We will not consider the extensive work on DCNNs for specific object detection, such as face detection, pedestrian detection , vehicle detection , and traffic sign detection, because our focus is on generic object detection.

presents the proposed work/experimental/simulation specifications:

As we have discussed that the aim of the project is to give a detailed analysis of the different deep learning techniques for generic object detection. We need to have some aspects on which the analysis should be done. We have considered six factors on which we analyze different deep learning techniques.

Dataset Name : There are four well-known datasets for general object detection: PASCAL VOC , ImageNet, MS COCO , and Open Images . The four datasets are the foundation for each detection challenge. Each challenge includes a publicly accessible image dataset, ground truth annotation, and standardized evaluation software, as well as an annual competition and workshop. VOC, COCO, ILSVRC, and Open Images detection datasets' most common object classes.

Detection Framework: As illustrated by the drastic shift from handcrafted features to learned DCNN features, there has been steady improvement in object feature representations and classifiers for recognition. In contrast, the basic "slide window" technique remains popular in terms of localization, though with some efforts to prevent exhaustive search. The number of windows, on the other hand, is vast and rises quadratically with the number of image pixels, and the necessity to search across several scales and aspect ratios expands the search space even more. As a result, designing efficient and effective detection frameworks is critical for lowering computing costs. the milestone approaches appearing since deep learning entered the field, organized into two main categories:

(a) Two stage detection frameworks, which include a preprocessing step for generating object proposals;
(b) One stage detection frameworks, or region proposal free frameworks, having a single proposed method which does not separate the process of the detection proposal

Learning Method: Learning method is an approach or algorithm used to gain practical knowledge about the techniques. The mostly used learning method is SGD - Stochastic Gradient Descent. Stochastic Gradient Descent (SGD) is a quick and easy way to fit linear classifiers and regressors to convex loss functions like (linear) Support Vector Machines and Logistic Regression.

Source Code: The source code plays a vital role of designing, developing And execution of different deep learning techniques. They help a developer to understand the problem , Scope of the input for the techniques and modeling of the data. Synchronized SGD ia also a type of learning method.

Backbone DCNN: In detection frameworks, CNN topologies function as network backbones. AlexNet, ZFNet, VGGNet, GoogLeNet, Inception series, ResNet, DenseNet, and SENet are examples of frameworks.
The tendency in architecture progression is towards increased depth: AlexNet has eight layers, VGGNet has sixteen, and more recently, ResNet and DenseNet have both reached the hundred-layer threshold. It was VGGNet and GoogLeNet who demonstrated that increasing depth can boost representational power. Because a substantial portion of the parameters come from the FC layers, networks like AlexNet, OverFeat, ZFNet, and VGGNet have a vast amount of parameters while being only a few layers deep. Although newer networks like Inception, ResNet, and DenseNet have a lot of depth, they have considerably less parameters because they don't require FC layers.

Detector Name: In this research analysis we included the following Detector Names for different techniques are SegDeep, DeepIDNet, MRCNN, ION, GBDNet, ACCNN, CPF, SMN, ORN, SIN,CoupleNet.

S No	Technique	About	Drawbacks	Application	Dataset name	Detection Framework	Learning Method	Source Code	Backbone DCNN	Detector Name
1	CNN	Convolutional Neural Networks (CNNs), the most representative models of deep learning, are able to exploit the basic properties underlying natural signals: translation invariance, local connectivity, and compositional hierarchies. A typical CNN, has a hierarchical structure and is composed of a number of layers to learn representations of data with multiple levels of abstraction. We begin with a convolution xl−1 ∗ wl (1) between an input feature map xl−1 at a feature map from previous layer l−1, convolved with a 2D convolutional kernel (or filter or weights) wl . This convolution appears over a sequence of layers, subject to a nonlinear operation σ, such that xl j = σ N l−1 i=1 xi−1 i ∗ wl i,j + bl j , (2) with a convolution now between the Nl−1	In particular, there is an extreme need for labeled training data and a requirement of expensive computing resources, and considerable skill and experience are still needed to select appropriate learning parameters and network architectures . Trained networks are poorly interpretable , there is a lack of robustness to degradations, and many DCNNs have shown serious vulnerability to attacks, all of which currently limit the use of DCNNs in	Face detection, Facial emotion recognition, Autonomous cars, object detection,auto translation, Cancer detection	ImageNet	Region Based	SGD	-	-	-

		input feature maps x^{l-1}_i and the corresponding kernel $w^l_{i,j}$, plus a bias term b^l_j . The elementwise nonlinear function σ (·) is typically a rectified linear unit (ReLU) for each element, σ (x) = max{x, 0}. (3) Finally, pooling corresponds to the downsampling/upsampling of feature maps. These three operations (convolution, nonlinearity, pooling) ; CNNs having a large number of layers, a “deep” network, are referred to as Deep CNNs (DCNNs), with a typical DCNN architecture.	real-world applications.							
2	RCNN	1.Region proposal computation Class agnostic region proposals, which are candidate regions that might contain objects, are obtained via a selective search . 2. CNN model fine tuning Region proposals, which are cropped from the image and warped into the same size, are used as the input for fine-tuning a CNN model pretrained using a large-scale dataset such as ImageNet. At this stage, all region proposals with ≥ 0.5 IOU 6 overlap with a ground truth box are	Multistage pipeline of sequentially-trained (External RP computation, CNN finetuning, each warped RP passing through CNN, SVM and BBR training); Training is expensive in space and time; Testing is slow	Autonomous driving, Facial Recognition,Smart Surveillance systems	-	Region Based	SGD,BP	Caffe Matlab	AlexNet	SegDeepM DeepIDNet

		defined as positives for that ground truth box’s class and the rest as negatives. 3. Class specific SVM classifiers training A set of class specific linear SVM classifiers are trained using fixed length features extracted with CNN, replacing the softmax classifier learned by fine-tuning. For training SVM classifiers, positive examples are defined to be the ground truth boxes for each class. A region proposal with less than 0.3 IOU overlap with all ground truth instances of a class is negative for that class. Note that the positive and negative examples defined for training the SVM classifiers are different from those for fine-tuning the CNN. 4. Class specific bounding box regressor training Bounding box regression is learned for each object class with CNN features.								
3	SPP Net	SPPNet During testing, CNN feature extraction is the main bottleneck of the RCNN detection pipeline, which requires the	While SPPNet accelerates RCNN evaluation by orders of magnitude,	Classification and object detection	-	Region Based	SGD	Caffe Matlab	ZFNet	MRCNN

		extraction of CNN features from thousands of warped region proposals per image. As a result, He et al. introduced traditional spatial pyramid pooling (SPP) into CNN architectures. Since convolutional layers accept inputs of arbitrary sizes, the requirement of fixed sized images in CNNs is due only to the Fully Connected (FC) layers, therefore He et al. added an SPP layer on top of the last convolutional (CONV) layer to obtain features of fixed length for the FC layers. With this SPPNet, RCNN obtains a significant speedup without sacrificing any detection quality, because it only needs to run the convolutional layers once on the entire test image to generate fixed-length features for region proposals of arbitrary size.	it does not result in a comparable speedup of the detector training. Moreover, fine-tuning in SPPNet is unable to update the convolutional layers before the SPP layer, which limits the accuracy of very deep networks							
4	Fast CNN	Fast RCNN Girshick proposed	Most of the time taken	Object detection	-	Region Based	SGD	Caffe Python	AlexNet VGGM	ION GBDNet

		Fast RCNN that addresses some of the disadvantages of RCNN and SPPNet, while improving on their detection speed and quality. Fast RCNN enables end-to-end detector training by developing a streamlined training process that simultaneously learns a softmax classifier and class-specific bounding box regression, rather than separately training a softmax classifier, SVMs, and Bounding Box Regressors (BBRs) as in RCNN/SPPNet. Fast RCNN employs the idea of sharing the computation of convolution across region proposals, and adds a Region of Interest (RoI) pooling layer between the last CONV layer and the first FC layer to extract a fixed-length feature for each region proposal. Essentially, RoI pooling uses warping at the feature level to approximate warping at the	by Fast R-CNN during detection is a selective search region proposal generation algorithm. Hence, it is the bottleneck of this architecture which was dealt with in Faster R-CNN.						VGG16	ACCNN

		image level. The features after the RoI pooling layer are fed into a sequence of FC layers that finally branch into two sibling output layers: softmax probabilities for object category prediction, and class-specific bounding box regression offsets for proposal refinement. Compared to RCNN/SPPNet, Fast RCNN improves the efficiency considerably—typically 3 times faster in training and 10 times faster in testing. Thus there is higher detection quality, a single training process that updates all network layers, and no storage required for feature caching								
5	Faster CNN	f Faster RCNN, Lenc and Vedaldi challenged the role of region proposal generation methods such as selective search, studied the role of region proposal generation in CNN based detectors, and found that CNNs contain sufficient	One drawback of Faster R-CNN is that the RPN is trained where all anchors in the mini-batch, of size	Object detection	-	Region Based	SGD	Caffe Python/ Matlab	ZFnet VGG	CPF SMN ORN SIN

		geometric information for accurate object detection in the CONV rather than FC layers. They showed the possibility of building integrated, simpler, and faster object detectors that rely exclusively on CNNs, removing region proposal generation methods such as selective search	256, are extracted from a single image. Because all samples from a single image may be correlated (i.e. their features are similar), the network may take a lot of time until reaching convergence							
6	R-FCN	RFCN (Region based Fully Convolutional Network) While Faster RCNN is an order of magnitude faster than Fast RCNN, the fact that the region-wise sub-network still needs to be applied per RoI (several hundred RoIs per image) I to propose the RFCN detector which is fully	Training is not a streamlined process; Still falls short of real time	Object detection	-	Region Based	SGD	Caffe Matlab	RPN	CoupleNet

		convolutional (no hidden FC layers) with almost all computations shared over the entire image., RFCN differs from Faster RCNN only in the RoI sub-network. In Faster RCNN, the computation after the RoI pooling layer cannot be shared, so Dai et al.proposed using all CONV layers to construct a shared RoI sub-network, and RoI crops are taken from the last layer of CONV features prior to prediction. However, found that this naive design turns out to have considerably inferior detection accuracy, conjectured to be that deeper CONV layers are more sensitive to category semantics, and less sensitive to translation, whereas object detection needs localization representations that respect translation invariance. Based on this observation, Dai et al constructed a set of position sensitive score								

		maps by using a bank of specialized CONV layers as the FCN output, on top of which a position-sensitive RoI pooling layer is added. They showed that RFCN with ResNet101 could achieve comparable accuracy to Faster RCNN, often at faster running times.								
7	Mask R-CNN	Mask RCNN He et al. (2017) proposed Mask RCNN to tackle pixelwise object instance segmentation by extending Faster RCNN. Mask RCNN adopts the same two stage pipeline, with an identical first stage (RPN), but in the second stage, in parallel to predicting the class and box offset, Mask RCNN adds a branch which outputs a binary mask for each RoI. The new branch is a Fully Convolutional Network (FCN) on top of a CNN feature map. In order to avoid the misalignments caused by the original RoI	Falls short of real time applications	satellite imagery, autonomous vehicles and medical applications, such as tumor detection or even detecting features related to the coronavirus.	-	Region Based	SGD	Tensor Flow - Python	ResNet101 ResNeXt101	-

		pooling (RoIPool) layer, a RoIAlign layer was proposed to preserve the pixel level spatial correspondence. With a backbone network ResNeXt101-FPN , Mask RCNN achieved top results for the COCO object instance segmentation and bounding box object detection. It is simple to train, generalizes well, and adds only a small overhead to Faster RCNN, running at 5 FPS								
8	FPN	FPN is built on utilising deep CNN's inherent multi-scale pyramidal hierarchy. It's similar to the difference between RCNN and Fast RCNN in that RCNN is a region-based object detector in which we first find ROI's using an algorithm like selective search and then crop these ROI's (around 2000) from the image and feed them into CNN to get results, whereas in Fast RCNN, the initial	-	Object detection	-	-	Synchronized SGD	Tensor Flow Python	-	-

		layers of CNN are shared for the entire image and the ROI cropping is done on the extracted feature map, saving a lot of time. The picture pyramid is somehow applied internally to architecture and sharing most sections of the network in the case of FPN, and the research is based on utilising internal multi-scale nature.								
9	YOLO	First efficient unified detector; Drop RP process completely; Elegant and efficient detection framework; Significantly faster than previous detectors; YOLO runs at 45 FPS, Fast YOLO at 155 FPS;	Accuracy falls far behind state of the art detectors; Struggle to localize small object	Object detection	-	Unified	SGD	Darknet - C	GoogLeNet like	-
10	YOLOv2	Propose a faster DarkNet19; Use a number of existing strategies to improve both speed and accuracy; Achieve high accuracy and high speed; YOLO9000 can detect over 9000 object categories in real time	Not good at detecting small objects	Object detection	-	Unified	SGD	Darknet - C	DarkNet	-
11	Over	Convolutional	Multi-stage	Object	-	Unified	-	-	AlexNet	-

	feat	feature sharing; Multiscale image pyramid CNN feature extraction; Won the ISLVRC2013 localization competition; Significantly faster than RCNN	pipeline sequentially trained; Single bounding box regressor; Cannot handle multiple object instances of the same class; Too slow for real time applications YOLO − GoogLeNet like Fixed 66.4 (07+12) 57.9 (07++12) < 25 (VGG) CVPR16 DarkNet	detection Classifiacation					like	
12	SSD	First accurate and efficient unified detector; Effectively combine ideas from RPN and YOLO to perform detection at multi-scale CONV layers; Faster and significantly more accurate than YOLO; Can run at 59 FPS;	Not good at detecting small objects	Object detection	-	Unified	SGD	Caffe - C++	VGG16	-

Conclusions and Future work.

Deep learning-based object detection has been a research hotspot in recent years due to its tremendous learning capacity and advantages in dealing with occlusion, size transformation, and backdrop shifts. This study gives a comprehensive overview of deep learning-based object detection frameworks that address a variety of subproblems, including occlusion, clutter, and low resolution, with varying degrees of R-CNN modification. The review begins with generic object identification pipelines, which serve as the foundation for other related activities. The core concept of Generic Object Detection is well presented, as are key elements such as Datasets, Detection Framework, Learning method, Source Code, Backbone DCNN and Detector Name for various detection strategies. Finally, to acquire a complete grasp of the object detection landscape, we offer various possible future possibilities. This review is especially useful for improvements in neural networks and related learning systems, since it offers useful insights and directions for future progress.

As, there is huge demand and also being a very developing topic, many changes can be made such as adding features to it and if possible trying to find other various techniques along with its characteristics would be good enough to perform future research.

Citations:

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, *128*(2), 261-318.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, *30*(11), 3212-3232.

Zhang, X., Yang, Y., Han, Z., Wang, H., & Gao, C. (2013). Object class detection: A survey. ACM Computing Surveys, 46(1), 10:1–10:53

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012a). ImageNet classification with deep convolutional neural networks. In NIPS (pp. 1097–1105).

.Brownlee, J., 2022. *A Gentle Introduction to Object Recognition With Deep Learning*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/object-recognition-with-deep-learning/> [Accessed 9 May 2022].