IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Object Detection using Deep Learning

Naman Joshi
(20EAICS1010)
Department of Computer Science engineering
Arya College of Engineering
Jaipur, India

Muskan Arora
(20EAIAD015)
Department of Artificial Intelligence and Data Science
Arya College of Engineering
Jaipur, India

***Abstract*— In recent years, there has been a lot of research focused on the difficult job of object detection in computer vision. Customised features and shallow trainable architectures, which are the foundation of traditional object identification techniques, are prone to performance stagnation. Conversely, deep learning techniques have greater strength and are able to learn deeper, semantic characteristics at a higher level. In addition to reviewing the background of deep learning and convolutional neural networks (CNNs), this study offers an overview of object identification frameworks based on deep learning. Typical general object detection structures, performance-enhancing changes, particular detection tasks, and experimental evaluations are also covered. Lastly, it makes encouraging recommendations for further research. To put it briefly, deep learning is an effective instrument for object detection that can outperform more conventional techniques. I strongly advise engineering students who are interested in object detection to read this paper in order to become more knowledgeable about the most recent methods.**

Keywords*— *object detection, deep learning, convolutional neural network

I. Introduction

For a thorough understanding of images, object detection is necessary for both object categorization and object location. It entails accurately recognising item ideas and whereabouts in pictures. This job contributes to several applications including picture classification, behaviour analysis, face recognition, and autonomous driving. It comprises subtasks like face, pedestrian, and skeletal detection. The ongoing development of neural networks has a major influence on object identification methods. Even with obstacles such different perspectives, positions, occlusions, and illumination, object identification is still a primary area of study. Three steps are usually included in traditional object detection models: feature extraction, classification, and informative region selection. This procedure is essential for identifying the locations and classes of objects in photos, which helps with semantic interpretation and opens up a number of useful applications.

Because different objects have unknown locations, aspect ratios, and sizes, it is sense to use a multi-scale sliding window to scan the whole image throughout the informative region selection process. Although this thorough approach may be able to pinpoint every location of an object, it has flaws. High computational costs and a large number of redundant windows are produced by the sheer amount of candidate windows. On the other hand, using a certain quantity of sliding window templates could result in insufficient coverage of pertinent areas. Thus, it is crucial to achieve a balance between template-based and comprehensive scanning methods in order to effectively find informative areas in object detection tasks.

The goal of the feature extraction step is to provide a robust and semantic representation by extracting visual properties that are necessary for differentiating between objects. SIFT, HOG, and Haar-like features are representative approaches that were selected because they can provide representations that resemble complicated brain cells. However, the difficulty stems from the many looks, lighting fluctuations, and backdrops that one encounters in real-world situations. Because of these difficulties, designing a feature descriptor that is reliable enough to manually capture all object kinds is difficult. As a result, developing a thorough and flexible feature representation for object identification is still a major goal in the discipline.

A classifier is also required to improve the hierarchical, semantic, and informative quality of representations for visual recognition, as well as to distinguish a target item from other categories. Generally, the best choices are Supported Vector Machine (SVM), AdaBoost, and Deformable Part-based Model (DPM). DPM is noteworthy for being a versatile model that successfully addresses severe deformations by integrating object components with deformation costs. To combine carefully designed low-level characteristics with component decompositions inspired by kinematics, DPM uses a graphical model. By discriminative learning of graphical models, it is possible to build high-precision part-based models customised for different classes of objects. Therefore, choosing the right classifier is essential to improving object detection systems' reliability and accuracy.

II. A ShortGuide To Deep Learning

Before providing an overview of deep learning-based object identification techniques, we first go over the background of deep learning and briefly discuss the fundamental design and benefits of CNN.

The History

Deep models—a term for neural networks with complex architectures—have their roots in the 1940s and are motivated by the goal of reproducing the structure and functions of the human brain through a methodical approach to a variety of learning tasks. Neural networks gained popularity in the 1980s and 1990s, particularly after Hinton et al. introduced the back-propagation algorithm. However, in the early 2000s, these networks encountered difficulties because of overfitting, a lack of training data, computational limitations, and inadequate performance in comparison. However, about 2006, advances in voice recognition led to a rebirth of deep learning. This revival was fueled by a number of reasons, including:

- The availability of extensive datasets with annotations, like as ImageNet, which demonstrate the remarkable ability of deep models to learn. Quick developments in GPU clusters and other high-performance parallel computing platforms.
- Notable advancements in the design of network architectures and training methodologies, such as batch normalisation (BN) for increased efficiency and dropout and layerwise and unsupervised pre-training with Auto-Encoders (AE) or Restricted Boltzmann Machines (RBM).
- Thorough investigation of many network architectures, including ResNet, VGG, GoogLeNet, Overfeat, AlexNet, and more, with the goal of improving performance.

Deep learning's significant influence on academia is a result of Hinton's group's innovative research. Their tireless efforts showed that deep learning might not just provide small-scale, incremental gains on tiny datasets, but potentially transform the way large-scale problems are solved. Their achievements, such as training massive CNNs on 1.2 million labelled pictures and using cutting-edge methods like dropout regularisation and ReLU activation, drove the use and development of deep learning approaches.

III. Generic Object Detection

In generic object detection, items in an image are found and categorised. Typically, rectangular bounding boxes that indicate confidence levels are used to indicate the locations of objects. Generic object detection techniques are broadly classified into two frameworks. To begin with, conventional methods work in a step-by-step pipeline, producing region suggestions and then categorising them into different item classes. On the other hand, an alternative method considers object detection as a regression or classification problem, directly determining categories and locations using a single framework. Among the region proposal-based techniques are R-CNN, SPP-net, Fast R-CNN, Faster R-CNN, R-FCN, FPN, and Mask R-CNN. A few of these techniques have correlations with each other (for example, SPP-net is just R-CNN enhanced with an SPD layer). The methods that bridge the gap between these two pipelines are called anchors presented in Faster RCNN. Regression and classification-based approaches include MultiBox, AttentionNet, G-CNN, YOLO, SSD, YOLOv2, DSSD, and DSOD.

Region Proposal Based Framework

The two-step procedure used by the area proposal based framework, which is represented by Overfeat, is comparable to the attention mechanism seen in the human brain. It scans the whole scene coarsely at first, then zeroes in on areas of interest. Using the sliding window method, Overfeat incorporates Convolutional Neural Networks (CNN) to predict bounding boxes based on the locations of the highest feature maps. This method finds the confidences of object categories effectively. It provides a simplified approach to object recognition, which is crucial for computer vision applications.

In response to the demand for improved candidate bounding boxes, Ross Girshick proposed R-CNN (Region-based Convolutional Neural Network) in 2014. This neural network uses a deep architecture to extract high-level information. With a mean average precision (mAP) of 53.3%, this methodology represents a notable improvement over earlier methods such as DPM HSC on the PASCAL VOC 2012 dataset. The three stages of the R-CNN flowchart, which is shown in Figure 3, all work together to increase the accuracy of object detection.

Region Proposal Generation: R-CNN generates about two thousand region ideas for each image by using selective search. This method quickly generates correct candidate boxes of different sizes by using saliency cues and basic bottom-up grouping. Selective search improves the speed and accuracy of locating regions of interest inside images by effectively decreasing the search space in object detection.

CNN based deep feature extraction : Every region suggestion is transformed or cropped in this phase in order to produce a resolution that is consistent. The final representation is then extracted as a 4096-dimensional feature using the CNN module from reference [6]. Using CNNs' expressive dominance, large learning capacity, and hierarchical architecture, this technique produces a high-level, semantic, and robust feature representation for every area suggestion.

Classification and localization: Different region recommendations are scored against positive and negative areas using pre-trained category-specific linear Support Vector Machines (SVMs) for several classes. The scored regions are then adjusted using bounding box regression and filtered using a greedy non-maximum suppression (NMS) technique. This process results in the final bounding boxes for the locations of the recognised objects. Pre-training is typically utilised in situations when there is a lack of labelled data. R-CNN first goes through supervised pre-training on ILSVRC, a large auxiliary dataset, as opposed to unsupervised pre-training, and then it is fine-tuned for a particular domain. R-CNN still has several shortcomings, despite its improvements over traditional techniques and its crucial role in incorporating CNN into useful object identification.

Classification and Localization : Different region recommendations are scored against positive and negative areas using pre-trained category-specific linear Support Vector Machines (SVMs) for several classes. The scored regions are then adjusted using bounding box regression and filtered using a greedy non-maximum suppression (NMS) technique. This process results in the final bounding boxes for the locations of the recognised objects. Pre-training is typically utilised in situations when there is a lack of labelled data. R-CNN first goes through supervised pre-training on ILSVRC, a large auxiliary dataset, as opposed to unsupervised pre-training, and then it is fine-tuned for a

particular domain. R-CNN still has several shortcomings, despite its improvements over traditional techniques and its crucial role in incorporating CNN into useful object identification.

IV. Salient Object Detection

Visual saliency detection is an important computer vision technique that enhances image cropping, segmentation, retrieval, and object detection by highlighting salient object locations in images. There are two types of approaches: top-down (TD) and bottom-up (BU). Low contrast saliency maps may result from BU techniques, which place an emphasis on local feature contrast at the expense of high-level semantic information. On the other hand, TD approaches use past object category information to direct the creation of salient maps, such as in semantic segmentation. By focusing attention on BU salient spots that are more likely to be components of the object, TD saliency serves as a focus-of-attention mechanism. This distinction emphasises how difficult it is to balance semantic relevance and local contrast in saliency detection, a task that is essential for improving computer vision applications.

A. Deep learning in Salient Object Detection

Expanding Convolutional Neural Networks (CNN) into salient object detection is a compelling need and feasible given its critical role in providing high-level, multi-scale feature representation and its success across various correlated computer vision tasks like edge detection, semantic segmentation, and generic object detection. This update promises improvements in recognising significant objects within photos by leveraging CNN's demonstrated effectiveness and versatility. Its use into saliency detection highlights CNN's potential to further transform computer vision applications and is consistent with the overall goal of improving visual processing skills.

V. Face Detection

Face detection is an essential preprocessing step for tasks like face recognition, synthesis, and facial expression analysis. It is used in many facial applications. Unlike generic object identification, it considers unique structural configurations (such the distribution of facial components) and attributes (like skin colour) of faces, and it includes detecting and locating face areas across a wide range of sizes (30-300 pixels vs. 10-1000 pixels). Large visual fluctuations, such as occlusions, position variations, and illumination changes, present a distinct challenge to the work and call for extra care. These intricacies highlight how important it is to develop face identification methods in order to successfully handle real-world problems.

Viola and Jones presented a well-known face detector that achieved excellent real-time efficiency by using cascaded classifiers using AdaBoost and Haar-like features. However, the high visual variety in human faces may cause this detector to significantly degrade in real-world circumstances. By contrast, Felzenszwalb et al. broke from the cascade structure and suggested a deformable part model (DPM) for face detection. However, in order to achieve decent results, standard face identification systems require a significant amount of computational power and comprehensive annotations. Furthermore, shallow architecture and hand-crafted features significantly limit their performance.

A. Face Detection using Deep Learning

New developments in CNN-based face detection are expanding the realm of possibilities. By concurrently forecasting all sides, novel techniques such as the IoU loss function proposed by Yu et al. enhance bounding box accuracy. Deep Dense Face Detector by Farfade et al. addresses multi-view face detection without requiring landmark or pose information. By concentrating on particular facial characteristics, Yang et al.'s Local Parts Network tackles challenging situations including occlusions and position changes. With specialised sub-networks, their ScaleFace network performs exceptionally well at detecting a range of facial sizes. Hao et al.'s approach, on the other hand, uses scale distribution prediction to focus on pertinent regions, improving performance at a lower computing cost. Furthermore, for increased face detection accuracy, current generic frameworks such as Faster R-CNN are being modified. The future of face detection is being shaped by these joint efforts..

VI. Pedestrian Detection

Because pedestrian detection is closely related to human re-identification, robot navigation, and pedestrian tracking, it has attracted a lot of attention recently. Some researchers merged boosted decision trees with hand-crafted features for pedestrian identification before advances in DCNN-based approaches. In the meantime, explicit occlusion handling approaches and part-based models have been investigated to directly address deformation and occlusion. Specifically, implementing a ROI pooling layer in generic object detection pipelines may result in inferior features in applications like automated driving and intelligent surveillance, where small-sized pedestrian cases are frequently encountered. Furthermore, in contrast to generic object detection, pedestrian detection frequently relies on hard background instances to identify incorrect predictions, requiring distinct configurations and components for accurate recognition.

A. Pedestrain Detection using Deep Learning

DCNNs have shown impressive results in generic object detection; nevertheless, none have been able to surpass the best manual feature-based technique over a sustained period of time. This discrepancy continues even when part-based information and occlusion handling are integrated. Scholars have carried out examinations to comprehend this occurrence. Zhang et al. modified the downstream classifier, added boosted forests to shared, high-resolution convolutional feature maps, and used a Region Proposal Network (RPN) for small instances and challenging negative examples in an effort to modify generic Faster R-CNN for pedestrian detection. Tian et al. proposed DeepParts, a DPM-inspired deep learning framework that uses an ensemble of extensive part detectors to address complex occlusions. DeepParts performs exceptionally well with partial blockage, low IoU positive proposals, and poorly labelled data.

VII. Scope Of Project

Even with the quick development and encouraging advances in object detection, there are still many open problems that need to be investigated in the future.

The initial concern pertains to the detection of small objects, commonly observed in datasets like COCO and in tasks such as face detection. Enhancing the accuracy of localization for small objects, especially under partial occlusions, requires architectural modifications to the network.

VIII. CONCLUSION

The field of deep learning-based object recognition has gained significant attention in recent years because of its strong learning skills and ability to handle various obstacles such as size transformation, background fluctuations, and occlusion. This study provides an extensive overview of deep learning-based object detection frameworks that address multiple sub-problems, including poor resolution, clutter, and occlusion, by varying the R-CNN architecture. Starting with a summary of generic object detection pipelines that function as foundational structures, the article goes on to address three typical tasks: face detection, pedestrian detection, and conspicuous object detection. In conclusion, we suggest multiple encouraging avenues for further investigation to further our comprehension of object detection techniques. This review is important for the development of neural networks and associated learning systems since it provides useful information and direction for future achivements.

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.