R-CNN for Object Detection

Sreeman Reddy K
Department of EP
IIT Bombay
sreeman@iitb.ac.in

Jahnavi Devangula
Department of EE
IIT Bombay
20d070025@iitb.ac.in

Kshitiz Susawat
Department of EE
IIT Bombay
19D070030@iitb.ac.in

Abstract

Region-based Convolutional Neural Network(R-CNN) is a deep learning architecture that achieves a good state-of-the-art performance on a wide range of object detection activities. One of the key innovation related to R-CNN is its better ability to perform object detection tasks by proposing Regions of Interest(RoI) in an image. It then classifies these regions using a convolutional neural network. This paper presents a replication study of the advanced research paper ”Rich feature hierarchies for accurate object detection and semantic segmentation” by Ross Girshick et al., which introduced the Region-based Convolutional Neural Network (R-CNN) for object detection. In this, we aim to reproduce the results that are reported in the original paper and evaluate the performance of R-CNN on several benchmark datasets. Our replication study aims to provide a deeper understanding of the R-CNN model and its performance on different datasets. We hope that our work will contribute in establishing best practices for evaluating and comparing several object-detection algorithms.

1. Introduction

Object detection is one of the important tasks in computer vision, with its wide range of applications such as robotics, autonomous driving and many more. One of the preferred approaches for object detection is the Region-Based Convolutional Neural Network (R-CNN). R-CNN is a pioneer which had introduced the idea of proposals of region. This has significantly improved the accuracy of the object detection when compared to already existing methods. Regional Convolutional Neural Networks has become a crucial foundation in the task of object detection. It has inspired several followed by research studies in the works of deep learning in object detection. In this, we will provide an overview of the R-CNN approach for object detection, including its architecture, training process, and its variations, and highlight its significant contributions to the field of computer vision.

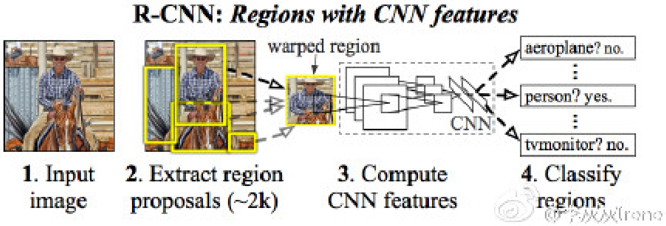

Figure 1. RoI in CNN [3]

2. Background and Prior Work

Object detection is a fundamental task that involves identification and localization of objects within a given image. In the recent years, the development of Deep Learning(DL) techniques have paved a way for prominent advances in this field of object detection, with several methods such as Region-based Convolutional Neural Networks(R-CNN) which has achieved the state-of-the-art performance on benchmark datasets. R-CNN is a type of 'multi-stage' object detection framework [3]. The time when R-CNN was not yet developed, object detection was basically performed using 'sliding-window' approaches or variants of the Viola-Jones algorithm [1]. One thing is to note that these methods were limited in their capability in handling the variations in object appearance. It often required a large number of hand-crafted features [2]. Introduction of deep learning techniques have enabled the development of more sophisticated models that could learn representations of objects directly from raw pixels. One of the earliest deep learning approaches to object detection was the OverFeat framework, which used a sliding window approach with a convolutional neural network (CNN) to classify and localize objects. While OverFeat achieved good performance, it was computationally expensive due to the need to evaluate the CNN at multiple scales.

R-CNN was proposed as an alternative to sliding window approaches that could handle object proposals more efficiently. The first stage of R-CNN involves generating a set of object proposals using an external algorithm, such as Selective Search. These proposals are then warped to a fixed size and passed through a pre-trained CNN to extract features. Finally, a set of SVM classifiers are trained to classify each proposal as either background or one of the object categories of interest, and bounding box regressors are trained to refine the proposal locations.

Since the original proposal of R-CNN, there have been several extensions and variants proposed, such as Fast R-CNN, Faster R-CNN, and Mask R-CNN, which have improved upon the original framework in terms of both accuracy and speed.

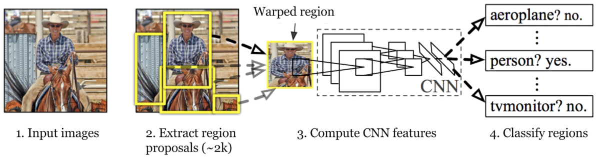

Figure 2. R-CNN [4]

3. Our Approach

R-CNN was proved to be one of the best deep learning techniques in image detection with its well-defined architecture. The important parts of this architecture are Region Proposal Network(RPN), Feature Extraction, Region of Interest(RoI) and image detection. The first stage of the R-CNN pipeline is to generate region proposals in which candidate bounding boxes that might contain objects of interest. R-CNN uses a selective search algorithm to generate around 2,000 region proposals per image [3]. The second stage involves extracting a fixed-length feature vector from each region proposal using a pre-trained convolutional neural network (CNN), such as AlexNet or VGGNet. The CNN is fine-tuned on a large-scale image classification dataset (such as ImageNet) to extract high-level features that are useful for detecting objects [3]. In the third stage, the feature vectors are fed into a set of class-specific linear Support Vector Machines (SVMs) to classify the region proposals and predict their bounding boxes. Each SVM is trained to distinguish between one object class and the background. The predicted bounding boxes are refined using a linear regression model that corrects the inaccuracies in the initial proposals [3]. R-CNN has been shown to achieve state-of-the-art performance on several benchmark object detection datasets, such as PASCAL VOC and MS COCO. However, the R-CNN architecture has several limitations, including slow inference speed due to the need to extract features separately for each region proposal and the inability to share computation across proposals. To address these limitations, several variants of R-CNN have been proposed, such as Fast R-CNN, Faster R-CNN, and Mask R-CNN, which build upon the original R-CNN architecture and improve its speed and accuracy [3]. The datasets used are PASCAL and VOC datasets. As part of comparision, lets compare three models R-CNN, YOLO (You Only Look Once), and SSD (Single Shot Detector). While all three methods use CNNs to detect objects, they differ in their approach to object detection and performance. R-CNN detects objects with excellent accuracy and is able to detect items of different sizes and forms. R-CNN can be customised on particular datasets, making it appropriate for particular applications. Yolo is suitable for applications like video surveillance since it is quick and can process photos in real-time. YOLO can find objects of all sizes and shapes and has a high detection rate. SSDs process photos quickly and in real time. Small objects and objects with different aspect ratios can be reliably detected by SSD. But Due to the requirement to extract features separately for each region proposal, R-CNN has a sluggish inference time. The feature vectors for each region suggestion must be stored in a significant amount of memory, which makes R-CNN another memory-intensive algorithm. Regarding the other two, because YOLO performs object detection at a coarse scale, it may miss small objects or objects with low contrast. Additionally, partially obscured or sculpted objects may be challenging for YOLO to identify.Large fluctuations in item size or aspect ratio may make it harder for SSD to recognise them. SSD can have trouble identifying things that are only partially obscured [3].

4. Experiments and results

A sizable auxiliary dataset was discriminatively pre-trained the CNN using only image-level annotations (boundingbox labels are not available for this data). Briefly stated, model in this paper achieved a top-1 error rate on the validation set that was 2.2 percentage points greater than Krizhevsky et al.'s performance [3]. The training method was simplified, which is the cause of this disparity. This was regarded as supervised pre-training. The stochastic gradient descent (SGD) training of the CNN parameters was continued using just warped area proposals in order to adapt CNN to the new goal (detection) and the new domain (warped proposal windows) [3]. The CNN architecture remains the same, same from replacing the ImageNet-specific 1000-way classification layer with a randomly initialised (N + 1)-way classification layer (where N is the number of object classes plus 1 for background). N is 20 for VOC dataset and 200 for ILSVRC2013 dataset [3]. For the class of a ground-truth box, we consider all region suggestions that have 3 0.5 IoU overlap with that box to be positive and the remainder to be negative. We begin SGD with a learning rate of 0.001 (1/10th of the initial pre-training rate), allowing fine-tuning to advance without completely destroying the system [3]. In each SGD iteration, we uniformly sample 32 positive windows (over all classes) and 96 background windows to construct a mini-batch of size 128. We bias the sampling towards positive windows because they are extremely rare compared to background [3]. This was part of Domain-specific fine-tuning. Considering the development of a binary

vehicle classification software. It is obvious that a narrowly constrained image region should serve as a good example. Likewise, it is evident that a background zone with no content

should serve as a bad example and relate to automobiles. How to designate a zone that partially encircles a car is less obvious. With an IoU overlap threshold, below which regions are classified as negatives, we solve this problem. In a grid search over 0–0.5 on a validation set, the overlap threshold of 0.3 was chosen [3]. We discovered that it's critical to properly choose this threshold. Its value of 0.5 caused mAP to drop by 5 points. Similarly, lowering it to 0 resulted in a 4-point drop in mAP [3]. Once features are extracted and training labels are applied, one linear SVM per class is optimized. Since the training data is too large to fit in memory, the standard hard negative mining method is adopted. Hard negative mining converges quickly and in practice mAP stops increasing after only a single pass over all images.

5. Discussion and Conclusion

On a number of benchmark object detection datasets, including PASCAL VOC and MS COCO, R-CNN has demonstrated state-of-the-art performance [3]. The R-CNN architecture, however, has a number of drawbacks, including sluggish inference speed because each region proposal requires that features be extracted independently and the inability to share compute between proposals. Several R-CNN variants, including Fast R-CNN, Faster R-CNN, and Mask R-CNN, which build on the original R-CNN architecture and enhance its speed and accuracy, have been proposed to address these limitations [3]. While YOLO and SSD are quick and suitable for real-time applications, they may not be as accurate in detecting some objects as R-CNN, which also has a slow inference speed. The selection of an object detection technique is based on the particular needs of the application. Utilising pre-trained CNN, R-CNN (Region-based Convolutional Neural Network) extracts features after using selective search to generate region proposals. Support vector machines (SVMs) are then fed these features in order to categorise the object inside each area proposed [3].

According to the paper's findings, R-CNN performs better than other cutting-edge object detection techniques on the PASCAL VOC 2012 and ILSVRC 2013 datasets. In comparison to its previous best score of 31.4, R-CNN's mAP score on PASCAL VOC 2012 is 53.3, which is a huge improvement. R-CNN significantly outperforms its previous high score of 15.8 on the ILSVRC 2013, achieving a top-5 accuracy of 31.4. By using it with the fully convolutional network (FCN) framework, the authors also demonstrate how R-CNN may be utilised for semantic segmentation. With a mIOU score of 53.3, the resulting technique, known as R-CNN FCN, performs at the cutting edge on the PASCAL VOC 2012 dataset [3].

The paper concludes by presenting a highly efficient method for semantic segmentation and object recognition utilising CNNs and region recommendations. The findings show that R-CNN outperforms other cutting-edge techniques, making it an important advancement in the field of computer vision.

6. Key Links

References

[1] Brown University. CS143: Introduction to Computer Vision - Project 4. https://cs.brown.edu/courses/cs143/2013/proj4/, 2013. Accessed on April 29, 2023. 1

[2] Data Science Stack Exchange. What is the meaning of ”hand-crafted features” in computer vision problems? https://datascience.stackexchange.com/questions/22782/what-is-the-meaning-of-hand-crafted-features-in-computer-vision-problems, 2018. Accessed on April 29, 2023. 1

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 1, 2, 3

[4] Lilian Weng. Object Recognition, Part 3: R-CNN Family. https://lilianweng.github.io/posts/2017-12-31-object-recognition-part-3/, 2017. Accessed on April 29, 2023. 2