Semiconductor Optoelectronics
ISSN:
1001-5868

AN EXPERIMENTAL ANALYSIS ON OBJECT DETECTION MODELS WITH DEEP LEARNING ALGORITHMS

Dr. Florence Vijila S
HOD of Computer Science, CSI Ewart Women's Christian College, Melrosapuram,
Chengalpet District, Tamil Nadu, India, florencevijila@yahoo.com

Abstract

For the past two decades, object detection has been a critical task and research focus in computer vision. It serves as the foundation for many inventions, including self-driving autonomous cars. A large number of objects are predefined and categorised quickly and accurately in a given image. There are two major types of algorithms for training a model. The first is a single stage detection algorithm, and the second is a two stage detection algorithm. This paper goes into detail about both types of algorithms and their specifications. The public common datasets were then used for image detection, and the various representative algorithms were analysed and compared. Finally, the paper concluded with the algorithm that produced better results, as well as potential challenges for object detection.

Keywords: Object detection, single stage detection, two stage detection, CNN, deep learning.

I. INTRODUCTION

Object detection is a fundamental research direction in computer vision, deep learning, artificial intelligence, and other fields. It is a necessary step toward more complex computer vision tasks like object tracking, behavior analysis, event detection, and extract semantic understanding. Its goal is to find the object of interest in the image, correctly determine the category, and provide the bounding box for each object. It is widely used in vehicle automatic driving, video and image retrieval, intelligent video surveillance[1, 2], medical image analysis[3, 4], and other fields.

Traditional detection algorithms for manually extracting features consist of six steps: pre-processing, window sliding, feature extraction, feature selection, feature classification, and post-processing. Its main disadvantages are small data size, poor portability, lack of pertinence, high time complexity, window redundancy, lack of robustness to diversity changes, and good performance only in specific simple environments. In the year 2012, Krizhevsjy[4] and others proposed the AlexNet image classification model based on convolutional neural network (CNN).

They won the image classification competition of the image datasetImageNet[5] with a huge 11% accuracy advantage over the second place using traditional algorithms. Scholars who began their work with deep convolutional neural networks to complete the task of object detection proposed many excellent object detection algorithms. That can be divided into two categories: single stage object detection algorithms based on region proposal and two stage object detection algorithms based on regression.

I1. FRAMEWORK OF TWO-STAGE OBJECT DETECTION

1.1 R-CNN

Girshick proposed the R-CNN [6] algorithm in 2014 as the first real object detection model based on CNN algorithm. The improved R-CNN model has a mAP of 66%. As illustrated in Figure 1, the model first employs Selective Search to extract approximately 2000 region proposals for each image to be detected. The extracted proposals are then uniformly scaled to a fixed-length feature vector, and the extracted image features are fed into the SVM classifier for classification. Finally, a linear regression model is trained to perform the bounding box regression operation.

The R-CNN algorithm's accuracy was greatly improved when compared to traditional CNN, but the only drawback was the large amount of calculation with very low efficiency. Scaling the proposed region directly to a fixed length feature vector causes object distortion.

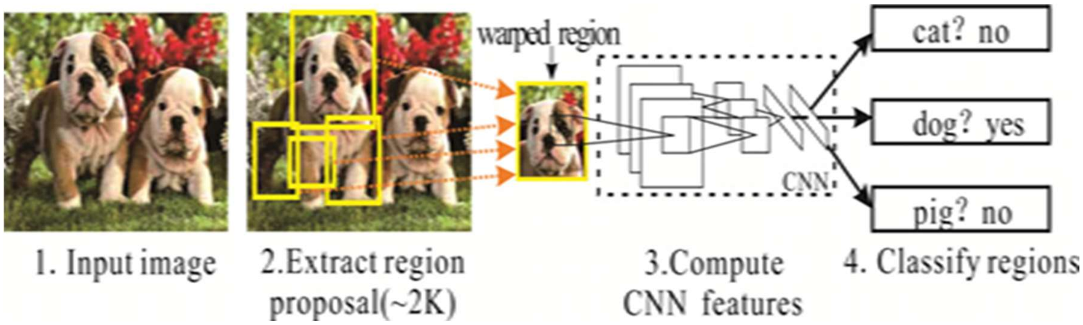

Figure 1: R-CNN Architecture

1.2 SPP-Net

He developed the SPP model in the year 2015. This model improves the R-CNN problem areas of low efficiency detection and fixed input size image blocks. After the original image is fed into the convolutional layer and all calculations are done in the convolutional layer, this algorithm extracts the proposed region on the feature map. At the end of the final convolutional layer, the special pyramid Pooling layer is added. The proposed region was fed into the SPP layer to generate the fixed size feature vector. SPP-Net outperforms R-CNN by avoiding repeated calculations and feature extraction on the entire image only once, despite having the same drawbacks. 1. Complications associated with multi-step training 2. The requirement for SVM classifiers and regressors.

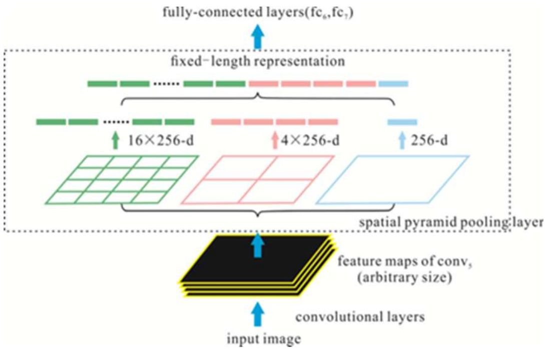
fully-connected layers(fc_6,fc_7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of $conv_5$
(arbitrary size)
convolutional layers
input image

Figure 2: SPP-Net Architecture

1.3 Fast R-CNN

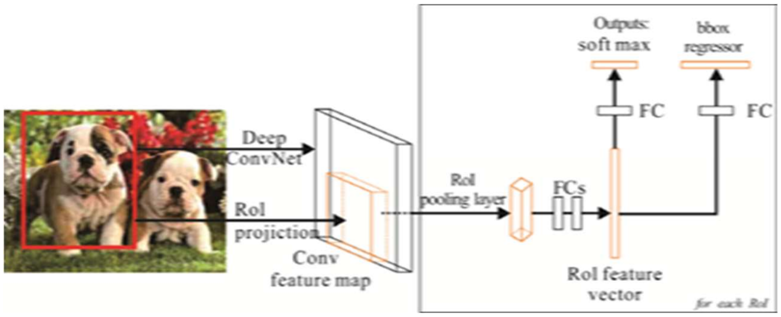

Figure 3: Fast R-CNN Architecture

Girshick proposed the Fast R-CNN[8] model in 2015. The mAP in the combined dataset of VOC2007 and VOC2012[15] is 70.0%. Figure 2 depicts its structure. When compared to R-CNN, Fast R-CNN has three differences. To begin, it replaced the SVM used in R-CNN with the softmax function for classification. Next, the model uses the pyramid pooling layer in SPP-Net, and the regions use the interest pooling layer to replace the last pooling layer in the convolutional layer, transforming the feature of the candidate box into a feature map with a fixed size for access to the entire connection layer. Finally, the CNN network's final softmax classification layer is replaced by two parallel fully connected layers. Even though new techniques have arrived, real-time detection is a moving target.

1.4 Faster R-CNN

Ren's Faster R-CNN[9] model replaces the previous Selective Search method for generating region proposals with region proposal networks. The model is divided into two modules: one is a fully convolutional neural network that is used to generate all region proposals, and the other is the Fast R-CNN detection algorithm. These two modules share a set of convolutional layers. The input image is passed through the CNN network until it reaches the final Shared convolutional layer. On the one hand, the RPN network's input feature map is obtained; on the other hand, the image is propagated forward to the specific convolutional layer to produce a higher-dimensional feature map. Even if Faster R-CNN excels at object detection accuracy, it cannot complete its process in real-time detection.

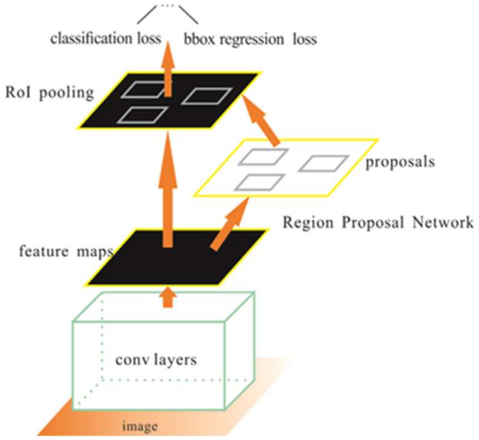

Figure 4: Faster R-CNN Architecture

2. ONE-STAGE OBJECT DETECTIONAL GEOMETRY

2.1 YOLOv1

The extraction of region proposals is not required in the YOLOv detection model. Joseph Redmon predicted this object detection model in the year 2016. The entire detection model is nothing more than a simple CNN network structure. Its central idea is to use the entire graph as the network's input and return the location and category of the bounding box directly at the output layer. First, an image is divided into S*S grid cells, with each grid cell predicting bounding boxes and confidence scores for these boxes. That is, each cell predicts a total of B*(4+1) values. On a single TitanX, detection speed can reach 45fps per second, allowing for fully real-time detection. YOLO, on the other hand, produces fewer background errors but has poor recognition performance when dealing with objects in groups.

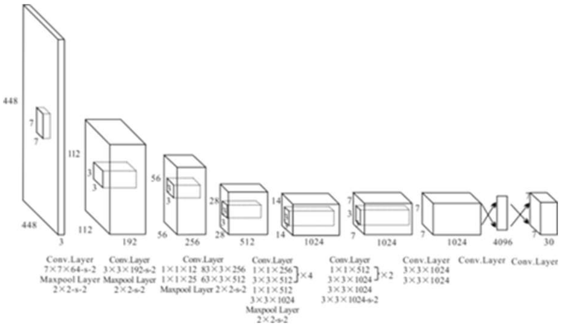

Figure 5: YOLOv1 Architecture

2.2 YOLOv2

Redmon proposed the YOLOv2[11] model in 2016. The main goal is to improve recall and localization while keeping classification accuracy constant. Darknet-19, a new fully convolution feature extraction network with 19 convolutional layers and 5 maximum pooling layers, is used in YOLOv2. The recall and accuracy are significantly improved by adding a batch normalisation layer to the convolutional layer and removing dropout, introducing an anchor box mechanism, using k-means clustering on the training set bounding box, and multi-scale training. However, detection of targets with high overlap and small targets requires further improvement.

2.3 YOLOv3

By far the most balanced object detection model for detection speed and detection accuracy is Redmon's YOLOv3[12]. In terms of category prediction, the main goal of YOLOv3 is to convert the original single-label classification into a multi-label classification, and to replace the original softmax layer used for single-label multi-classification with a logistic regression layer for multi-label multi-classification. Simultaneously, the model predicts using a combination of multiple scales. It uses an up sampling fusion method similar to FPN and finally merges three scales, which significantly improves the detection effect of small targets. This model's network structure is based on the Darknet-53 deeper feature extraction network. Although the YOLOv3 model significantly improves detection speed and the detection effect of small targets, it does not significantly improve detection accuracy, especially when IOU>0.5.

2.4 SSD

In the year 2016, Liu created the SSD model. The model incorporates the regression concept from the YOLO algorithm as well as the anchor box concept from the Faster R-CNN detection model. SSD model proposes using both bottom and high level feature maps for detection to improve the effect of multi-scale object detection. The basic architecture is VGG, with convolutional layers replacing the last two fully connected layers. SSD makes use of the RPN network's anchor mechanism. On VOC2007, SSD achieves 74.3% mAP at 59 FPS on aNvidia Titan X. However, the SSD classification result for small targets is poor, and the feature maps of different scales are independent, resulting in the detection of the same object by boxes of different sizes at the same time.

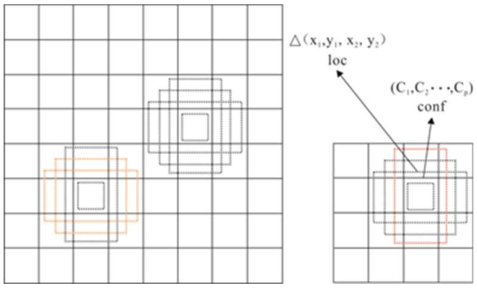

Figure 6: SSD Architecture

2.5 YOLOv4

Alexey Bochkovskiy proposed the YOLOv4[14] in 2020, and it achieves a new benchmark with the best balance of speed and accuracy. In theory, YOLOv4 is not particularly innovative. On the basis of the original YOLO detection framework, it adds Weighted Residual Connection, Cross Stage Partial Connection, Cross Mini Batch Normalization, Self-adversarial training, Mish activation, Mosaic data augmentation, DropBlock, and CIou. As the backbone network, CSP Darknet53 was chosen, and an SPP module was attached to increase the receptive field and separate the most important context features. Meanwhile, YOLOv4 employs PANet as the path aggregation method, rather than the FPN used in YOLOv3, and retains the YOLOv3 head structure. When compared to the YOLOv3, the YOLOv4 improves accuracy and speed by 10% and 20%, respectively.

III VARIOUS ALGORITHMS PERFORMANCE COMPARISON

3.1 Dataset

The term artificial intelligence (AI) was coined in 1956. But it wasn't until 2012 that artificial intelligence reached a tipping point. This is primarily due to increased data volume and computing power, as well as the emergence of machine learning algorithms. The evolution of detection systems is inextricably linked to the explosion of data volume. The dataset is essential in all object detection models. The dataset used determines the evolution of algorithm efficiency and accuracy.

The considerations of common public data sets are shown in table 1.

Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004
PASCAL VOC 2007	9963	20	375×500	2005
PASCAL VOC 2012	11540	20	470×380	2005
Tiny Images[19]	80 million	53464	32×32	2006
Scenes15	4485	15	256×256	2006
Caltech256	30607	256	300×200	2007
ImageNet	14197122	21841	500×400	2009
SUN[16]	131072	908	500×300	2010
MS COCO[17]	328000	91	640×480	2014
Places[20]	More than10 million	434	256×256	2014
Open Images	More than 9 million	More than 60 million	Different size	2017

TABLE I. PUBLIC DATA SET AND ITS PARAMETERS

3.2 Comparison of algorithm performance

Table 2 Comparisons of single-stage and two-stage detection algorithms.

Method	Backbone	Size/Pixel	Test	mAP/%	fps
YOLOv1	VGG16	448×448	VOC 2007	66.4	45

SSD	VGG16	300×300	VOC 2007	77.2	46
YOLOv2	Darknet-19	544×544	VOC 2007	78.6	40
YOLOv3	Darknet-53	608×608	MS COCO	33	51
YOLOv4	CSP Darknet-53	608×608	MS COCO	43.5	65.7
R-CNN	VGG16	1000×600	VOC2007	66	0.5
SPP-Net	ZF-5	1000×600	VOC2007	54.2	-
Fast R-CNN	VGG16	1000×600	VOC2007	70.0	7
Faster R-CNN	ResNet-101	1000×600	VOC2007	76.4	5

TABLE 2. COMPARISON OF OBJECT DETECTION ALGORITHMS

IV CONCLUSION

As one of the most basic and challenging problems in computer vision, object detection has received great attention in recent years. Detection algorithms based on deep learning have been widely applied in many fields, but deep learning still has some problems to be explored:

1) Reduce the dependence on data.
2) To achieve well-organized, efficient, accurate detection of small objects.
3) Realization of multi-category object detection.

REFERENCES

[1] Wu, R.B. Research on Application of Intelligent Video Surveillance and Face Recognition Technology in Prison Security. China Security Technology and Application. 2019,6:16-19.

[2] Tian, J.X., Liu, G.C., Gu, S.S., Ju, Z.J., Liu, J.G., Gu, D.D. Research and Challenge of Deep Learning Methods for Medical Image Analysis. ActaAutomatica Sinica,2018, 44:401-424.

[3] Jiang, S.Z., Bai, X. Research status and development trend of industrial robot target recognition and intelligent detection technology. Guangxi Journal of Light Industry, 2020, 36:65-66.

[4] Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems,2012, 25:1097-1105.

[5] Russakovsky, O., Deng, J., Su, H., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision,2015, 115:211-252.

[6] Girshick, R., Donahue, J., Darrel, T.,Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Computer Vision and Pattern Recognition. Columbus.2014, pp.580-587.

[7] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015, 37:1904-1916.

[8] Girshick, R. Fast R-CNN.In: Proceedings of the IEEE international conference on computer vision. Santiago.2015, pp.1440-1448.

[9] Ren, S.Q., He, K.M., Girshick, R., Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montreal.2016, pp.91-99.

[10] Redmon, J., Divvala, S., Grishick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In: Computer Vision and Pattern Recognition. Las Vegas.2016, pp. 779- 788.

[11] Redmon, J., Farhadi, A. YOLO9000: better, faster, stronger. In: Computer Vision and Pattern Recognition. Hawaii.2017, pp.7263-7271.

[12] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv: Computer Vision and PatternRecognition.

[13] Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, 2016, pp.21-37.

[14] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: Computer Vision and Pattern Recognition,2020.

[15] Everingham, M., Eslami, S.M.A., Van Gool, L. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision,2015,pp.98-136.

[16] Xiao, J.X., Ehinger, K.A., Hays, J.,Torralba, A.,Oliva, A. SUN Database: Exploring a Large Collection of Scene Categories. International Journal of Computer Vision,2016,pp.3-22.

[17] Lin T Y ,Maire M , Belongie S , et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision, 2014,pp.740-755.

[18] Li, F.F., Rob, F., Pietro, P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding,2007,pp.59-70.

[19] Torralba, A., Fergus, R., Freeman, W.T. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008,pp.1958-1970.

[20] Zhou, B., Lapedriza, A., Khosla, A., et al. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, pp.1452-1464.