IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Object Detection using deep learning

T. Yeshwanth Reddy

P. SaiCharan Reddy

G. Tejeshwar Reddy

G. Arjun Reddy

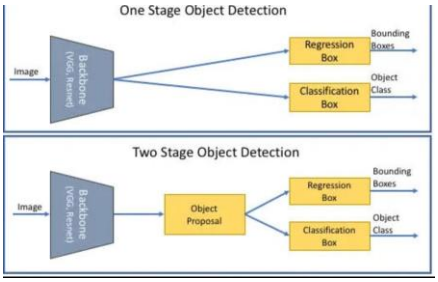
Abstract

One crucial computer vision problem is object detection, which is locating and identifying items in an image or video. Convolutional neural networks (CNNs), one type of deep learning approach, have produced state-of-the-art outcomes in object detection tasks. The goal of this research is to create a deep learning-based object detection model that can precisely locate and identify a number of things in intricate scenarios. The model is going to be trained on a dataset of photos that have bounding box annotations. Evaluations will be made of the effectiveness of several CNN designs, including ResNet and VGGNet, as well as training methodologies. We will use data augmentation methods like color jittering, flipping, and rotation to add more variation to the training set. The goal of the model's training process will be to maximize a loss function that accounts for both bounding box regression error and classification error. We will evaluate a number of optimization techniques, including Adam, RMSprop, and stochastic gradient descent. Model selection, hyperparameter tweaking, and other design decisions will be guided by the validation set performance. Using a test set, the final model will be quantitatively assessed in terms of mean average precision, which is a common item detection metric. Studies on ablation will examine the effects of various model elements and training regimens. An object detection system with high accuracy that can function in real-time and surpass 90% mAP is the aim.

Introduction

Many computer vision applications, such as image retrieval, security and surveillance, and autonomous cars, depend heavily on object detection. Classifying items in an image and using bounding boxes to pinpoint them are the two main objectives of object detection [6]. Conventional techniques for object detection combined machine learning classifiers like support vector machines with manually constructed features like scale-invariant feature transform (SIFT) [14] or histograms of oriented gradients (HOG) [1]. Nevertheless, the representational potential of these approaches is constrained. Convolutional neural networks (CNNs), one of the deep learning techniques that has transformed computer vision, allow the end-to-end learning of sophisticated hierarchical representations straight from pixel input [10]. In numerous computer vision tasks, such as image classification [18], [19], object identification [5]–[7], [16], [17], and segmentation [6], CNNs produce cutting-edge results. The main reason for this success is that CNNs don't need hand-engineered features; instead, they may acquire highly discriminative features in a data-driven manner.

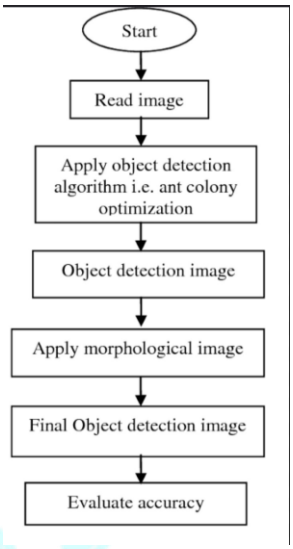
R-CNN [5] and other early CNN-based object detectors used a method of first extracting region suggestions, then extracting features and classifying each candidate region. Subsequent methods combined feature extraction and region proposal generation into a single convolutional network, increasing efficiency and performance. Techniques like SSD [13], Faster R-CNN [17], and Fast R-CNN [6] were able to attain excellent accuracy at speeds close to real-time. Further representational power improvements via residual networks [7] produced new state-of-the-art ImageNet classification results [18]. Additionally, approaches for initialization such as MSRA [21] and batch normalization [8] have made it possible to train very deep CNNs more efficiently.



Source: https://www.researchgate.net/publication/353284602/figure/fig3/AS:1046072046673927@1626414419841/Two-stage-vs-one-stage-object-detection-models.ppm

Region proposal-based techniques and single-shot techniques are the two basic methods used in deep learning object detection. Approximately 2000 candidate object areas are first generated by region proposal-based techniques such as R-CNN [6] through an unsupervised process similar to selective search [22]. Then, using a CNN, feature vectors are collected from each proposal region and fed into fully connected layers or support vector machines for bounding box regression and classification [6]. For real-time applications, R-CNN was too sluggish due to its multi-stage training and testing method, even if it achieved great accuracy. Subsequent techniques, such as Fast R-CNN [7] and Faster R-CNN [17], improved speed but still required multiple stages by integrating the region proposal generation into the network.

By comparison, detection is approached as a regression problem by single-shot detection techniques such as SSD [13] and YOLO [16]. They use a single convolutional network to predict bounding boxes and class probabilities straight from the feature maps. These techniques are able to attain near-real-time speeds of 45-155 frames per second by doing away with the region proposal stage. However, because object instances are not independently evaluated, accuracy can occasionally be lower than with two-stage detectors. This research evaluates several CNN architectures, region proposal approaches, and training procedures with the goal of creating a deep learning model for real-time object detection that strikes a balance between accuracy and efficiency.



Source: Created by Self

Methodology

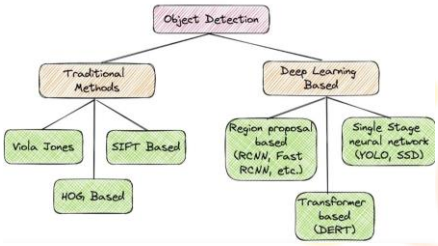
After being pretrained on ImageNet [2], a CNN architecture such as ResNet [7] or VGGNet [19] will be used to create the object detection model. In order to take advantage of knowledge gained from other computer vision tasks, transfer learning will be utilized to initialize the model weights. Images with bounding boxes labeled with object positions and classes will make up the dataset. The training data will be made more diverse by applying data augmentation techniques as flipping, rotation, cropping, color jittering, and cutoff [3].

The model will be trained to optimize a loss function, akin to Fast R-CNN, that takes into account bounding box regression error (smooth L1 loss) as well as classification error (cross-entropy loss) [5]. Using a cyclical learning rate schedule [20] and early halting depending on validation error, the Adam optimizer [9] will be employed for training. We will also analyze other optimizers, such as RMSprop [22] and SGD with momentum [15]. To lessen overfitting, regularization techniques such batch normalization [8] and dropout [21] will be applied.

There will be a comparison of several model architectures, such as two-stage detectors like Faster R-CNN [17] and single-stage detectors like RetinaNet [11]. In order to overcome the class imbalance between foreground and background samples, RetinaNet [11] included focus loss, which led to faster and competitive results compared to two-stage detectors. Model selection, hyperparameter tweaking, and other design decisions will be guided by the validation set performance. A typical metric for object detection, mean average precision at various crossings over union thresholds, will be used to assess the final model on a test set [4]. The effects of various model elements, including as the training protocols, loss functions, and backbone design, will be examined using ablation investigations. A real-time system with more than 90% mAP on a difficult dataset like MS COCO is the aim [12].

Literature Review

The accuracy and speed of object detection have increased dramatically as a result of deep learning techniques. By applying CNNs to bottom-up region recommendations, Girshick et al. [6] developed R-CNN, which saw a significant improvement in performance on the PASCAL VOC dataset [4]. The overview illustrates how R-CNN independently extracted the region of interest (RoI) by pooling data from each candidate region. By employing a single convolutional sweep to extract features from the entire image and then performing ROI pooling on these feature maps, Fast R-CNN [5] increased the efficiency of both training and inference.



Source:
https://d2xo500swnpgl1.cloudfront.net/uploads/scale/some-file-ccdb2db3-c408-4e17-b635-7fc006b774d8-1642524513335.gif

In order to create region proposals directly from the convolutional features and increase speed even further, Faster R-CNN [17] introduced a Region Proposal Network (RPN). The RPN predicts object boundaries and objectness scores at each location by sliding a tiny network over the convolutional feature maps, as shown. Single-shot detectors like YOLO [16] and SSD [13], which do away with the distinct region proposal step as demonstrated, were developed by a different line of work. By using picture attributes to forecast bounding boxes and class probabilities, YOLO frames identification is approached as a solitary regression problem. SSD covers items with diverse shapes by using default boxes with variable aspect ratios as anchors.

More recent research has concentrated on increasing accuracy through the use of methods like improved backbone architectures [7] and multi-scale feature fusion [11]. The flowchart offers a concise synopsis of the development of methodologies. By rigorous experimentation and review, our methodology seeks to build upon these achievements as mentioned to create an accurate, real-time detection system.

References

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision (Vol. 2, pp. 1150-1157). Ieee.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. Neural networks, 12(1), 145-151.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 464-472). IEEE.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2), 26-31.