A REVIEW OF RESEARCH ON DEEP LEARNING-BASED OBJECT DETECTION

Rasika Sachin Golhar
Research Scholar
Department of Computer Science & Engineering
Sunrise University, Alwar, Rajasthan.
rasikarode222@gmail.com

Dr. Pawan Kumar Pareek
Research Guide
Department of Computer Science & Engineering
Sunrise University, Alwar, Rajasthan.

Abstract

Target detection has been a significant research hotspot and an extensively utilized problem in computer vision during the last 20 years. In a given picture, it seeks to rapidly and precisely detect and locate a large number of items according to predetermined categories. The algorithms may be split into two categories based on the model training method: single-stage detection algorithms and two-stage detection algorithms. The typical algorithms for each level are thoroughly presented in this work. Then, numerous typical techniques are examined and contrasted in this area while public and special datasets that are often utilized in target identification are presented. The probable difficulties in target detection are therefore anticipated.

INTRODUCTION

In the disciplines of computer vision, deep learning, artificial intelligence, etc., object detection is a fundamental study area. For more difficult computer vision tasks like target tracking, event detection, behavior analysis, and scene semantic comprehension, it serves as a crucial precursor. It seeks to properly identify the category, find the target of interest inside the picture, and provide the bounding box of each target. It is often utilized in areas such as intelligent video surveillance, medical image analysis, automated driving of vehicles, industrial inspection, and video and picture retrieval.

Pre-processing, window sliding, feature extraction, feature selection, feature classification, and post-processing are the six basic processes in traditional detection techniques for manually extracting features, which are often used for specialized identification tasks. Small data size, low portability, lack of pertinence, high temporal complexity, window redundancy, lack of resilience for diversity changes, and strong performance only in certain basic situations are the primary drawbacks of this technology.

Krizhevsjy and colleagues introduced the AlexNet image categorization model based on convolutional neural network (CNN) in 2012.They defeated the second-place team utilizing conventional techniques by a large margin of 11% accuracy in the picture classification competition for the ImageNet image dataset. Deep convolutional neural networks have been used to target identification applications by various academics, who have also put out several great techniques. The single-stage detection method based on area proposal and the two-stage detection technique based on regression may be essentially classified into two groups.

TWO-STAGE TARGET DETECTION FRAMEWORK

R-CNN

The R-CNN method, the first practical target identification model based on convolutional neural networks, was put out by Girshick in 2014. The mAP for the enhanced R-CNN model is 66%.

Finally, a linear regression model is trained to carry out the bounding box regression process. The R-CNN does significantly enhance accuracy when compared to the conventional detection approach, but it requires a lot of calculations and does it inefficiently. Second, converting the region suggestion to a fixed-length feature vector directly could distort the objects.

SPP-Net

In 2015, He] developed the Spatial Pyramid Pooling (SPP) model to address the issues of R-CNN's poor detection performance and the need for fixed input size picture blocks. After the original picture has been processed by the convolution layer and just once via the convolution computations, this approach extracts the features of the areas proposed on the feature map. The spatial pyramid pooling layer is simultaneously added after the final convolutional layer, and the feature of the area proposal is passed through the layer to extract the feature vector with a defined size. Spp-Net only does feature extraction on the full picture once, as opposed to the R-CNN's recurring computations. However, it still shares R-CNN's drawbacks:

1) Complicated multi-step training procedures.

2) Additional regressors and separate SVM classifiers must be trained.

Fast R-CNN

Girshick put out the Fast R-CNN model in 2015. The mAP reaches 70.0% in the combined dataset of VOC2007 and VOC2012. Three modifications have been made to Fast R-CNN in comparison to R-CNN. First, it employed the softmax function for classification instead of the SVM that was used in R-CNN. In order to convert the feature of the candidate box into a feature map with a fixed size for access to the full connection layer, the model also draws on the pyramid pooling layer in SPP-Net and uses the region of interest pooling layer to replace the last pooling layer in the convolutional layer. Finally, two parallel fully linked layers are used to replace the CNN network's final softmax classification layer. But it still falls short of real-time detection requirements.

Faster R-CNN

Ren's Faster R-CNN model replaces the prior Selective Search approach for region proposal generation with region proposal networks. The model is composed of two modules: the Fast R-CNN detection method and a fully convolutional neural network used to produce all region proposals. These two modules share a set of convolutional layers. The input picture is sent via the CNN network to the shared convolutional layer at the very end. In order to create a higher-dimensional feature map, the picture is transmitted forward to the specified convolutional layer on the one hand, and the feature map for the input of the RPN network on the other. Even though Faster R-CNN has good detection accuracy, real-time detection is still not possible with it.

ONE-STAGE TARGET DETECTION ALGORITHM

YOLOv1

Joseph Redmon presented the YOLOv^1 object detection methodology in 2016. The extraction procedure of region proposals is not necessary for the YOLOv^1 detection model. Simply said, the whole detection model is a CNN network structure. The fundamental concept is to immediately return the position and category of the bounding box at the output layer by feeding the network the whole graph as input. An S*S grid is first used to break up the picture, and each grid cell predicts a B bounding box and the confidence scores for these boxes. In other words, each cell forecasts a total of B*(4+1) values. Its real-time detection rate on a single TitanX may be as fast as 45 frames per second. Although YOLO generates less background errors, it performs poorly when dealing with items that are grouped together.

YOLOv2

Redmon put out the YOLOv^2 concept in 2016. Enhancing recollection and localisation while keeping classification accuracy is the major objective. Darknet-19, a novel fully convolutional feature extraction network with a total of 19 convolutional layers and 5 maximum pooling layers, is the network used by YOLOv^2. The recall and accuracy are considerably increased by adding a batch normalization layer to the convolutional layer, eliminating dropout, implementing an anchor box method, applying k-means clustering on the training set bounding box, and multi-scale training. However, there is still room for improvement in the identification of targets with significant overlap and tiny targets.

YOLOv3

By far, Redmon's YOLOv^3 object detection model is the best balanced in terms of both detection speed and accuracy. YOLOv^3's primary goal in terms of category prediction is to convert the original single-label classification into a multi-label classification and swap out the original softmax layer with a logistic regression layer for multi-label multi-classification. The model makes predictions using a mixture of many scales at the same time. It uses a technique akin to FPN's upsampling fusion approach to combine three scales, greatly enhancing the identification of tiny objects. This model's network structure uses the deeper Darknet-53 feature extraction network. Although the YOLOv^3 model further increases detection speed and greatly increases the detection impact of tiny objects, it does not significantly increase detection accuracy, particularly when IOU>0.5.

SSD

Liu put out the SSD model in 2016. The model makes use of the YOLO algorithm's regression notion and the anchor box idea put out by the Faster R-CNN detection model. The SSD model suggests using both the bottom and top level feature maps for detection in order to enhance the effectiveness of multi-scale object detection. The last two fully linked layers are swapped out for convolution layers in the basic VGG architecture. SSD makes use of the RPN network's anchoring system. On an Nvidia Titan X, SSD scores 74.3% mAP on VOC2007 at 59 frames per second. However, the SSD's classification performance for tiny targets is subpar, and since the feature maps for different scales are independent, the same item may be simultaneously detected by boxes of various sizes.

YOLOv4

Alexey Bochkovskiy proposed the YOLOv^4 in 2020, and it sets a new standard with the best balance of speed and accuracy. Theoretically, YOLOv^4 isn't really inventive. On the foundation of the original YOLO detection system, it includes Weighted Residual Connection, Cross Stage Partial Connection, Cross small Batch Normalization, Self adversarial training, Mish activation, Mosaic data augmentation, DropBlock, and CIou. In order to broaden the receptive field and segregate the most crucial context characteristics, SPP module was coupled to CSP Darknet, which was chosen as the backbone network. The route aggregation mechanism in YOLOv[4] is PANet rather than the FPN utilized in YOLOv[3], and it retains the same head shape. The YOLOv[4] is 10% and 20% faster than the YOLOv[3] in terms of accuracy and speed, respectively.

DATASETS AND PERFORMANCE COMPARISON OF VARIOUS ALGORITHMS

Dataset

The idea of "artificial intelligence" was put up as early as 1956. However, it wasn't until 2012 that the rise of artificial intelligence really took off. The development of machine learning techniques, increased processing power, and growing data volumes are the key causes of this. The growth of data volume and the development of detecting technologies go hand in hand. This is due to the fact that datasets are required for performance testing and algorithm assessment and datasets are also a strong motivator for the advancement of the detection techniques study area.

CONCLUSION

Object identification is one of the most fundamental and difficult issues in computer vision and it have attracted a lot of attention lately. Although deep learning-based detection techniques have been extensively used in various domains, there are several issues that need to be investigated.

- Reduce the dependence on data.
- To achieve efficient detection of small objects.
- 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. Acta Automatica 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 Pattern Recognition.

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.