Literature review of YOLO Algorithm

DAI Guohao
Group 5B, 2A-SN, ENSEEIHT

Abstract

Object detection technology is the fundamental research in the field of artificial intelligence. Now it has strong applicability as well as been wildly used. This paper provides a brief overview of the current applications of object detection technology and an object detection algorithm named You Only Look Once (YOLO). Through comparison and analysis, we draw the differences and similarities between YOLO versions and between YOLO and another algorithm, Region-based Convolutional Neural Networks (R-CNN). The central insight is that the improvement of the YOLO algorithm is still in progress.

Background

A report released by the British Security Industry Association (BSIA) estimated that the total number of CCTV cameras in the UK as of now is between 4 million and 6 million. Furthermore, this number is 50 million in the United States and 200 million in China. According to "CCTV in the UK White Paper" published by British company *SYNECTICS*, "The results of the study revealed strong support for public space surveillance, with 86% of people backing its use, 76% think the primary purpose of CCTV should be to help prevent crime and anti-social behavior." [1] Object detection technology can be applied to CCTV equipment to detect the characteristics of body, clothing, hairstyle. So that it will quickly identify suspects from information.

Introduction

You Only Look Once (YOLO) is a widely used algorithm [2]. It is illustrious for its object detection characteristic. Since the initial version of the YOLO algorithm was introduced by the Redmon et al, scholars have published 5 iterations in order to make the algorithm more accurate and efficient.

Essential features of YOLO

The core of YOLO not only lies in the model's miniature size but also has superior computing efficiency. Its structure is straightforward so that can directly output the position and category of the bounding box through the neural network. Thanks to its rapid pace of execution, YOLO only needs to put the picture into the network to get the final detection result. Thus, it can further realize the real-time detection of video. The algorithm has a strong generalization ability by reason of YOLO can learn highly generalized features to be transferred to other fields. It converts the problem of target detection into a regression problem, but detection accuracy needs to be improved [3].

YOLO's test results are poor for objects that are very close to each other and in groups. This poor performance is because only two boxes in the grid are predicted and only belong to a new class of objects of the same category, as a result, an abnormal aspect ratio appears, moreover other conditions, such as weak generalization ability.

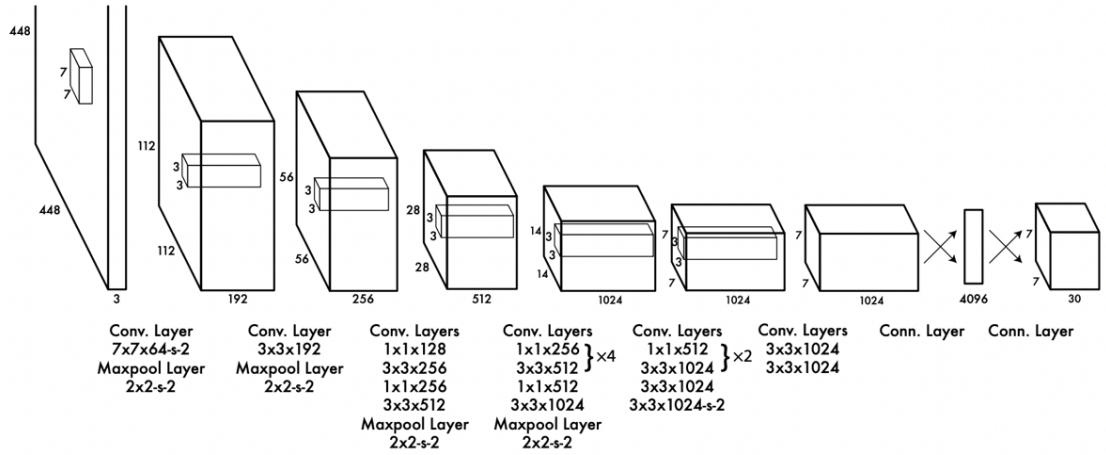

Figure 1 : Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers.

The original YOLO architecture consists of 24 convolution layers, followed by two fully connected layers. The algorithm predicts multiple bounding boxes per grid cell but those bounding boxes having highest Intersection Over Union (IOU) with the ground truth is selected, which is known as non-maxima suppression [4].

Nevertheless, it has two defects: one is inaccurate positioning, the other is the lower recall rate compared with the method based on area recommendations. Therefore, since the introduction of the YOLO algorithm, 5 versions have been iterated, besides, they are still being improved.

Main improvement measures of YOLO network from V1 to V5 [4][5][6]:

- YOLO V1: The grid division is responsible for detection, confidence loss.
- YOLO V2: Anchor with K-means added, two-stage training, full convolutional network.
- YOLO V3: Multi-scale detection by using FPN.
- YOLO V4: SPP, MISH activation function, data enhancement Mosaic/Mix-up, GIOU loss function.
- YOLO V5: Flexible control of model size, application of Hard swish activation function and data enhancement.

Comparison between YOLO and R-CNN

R-CNN and its variants use region proposals instead of sliding windows to find objects in images. Selective Search [7] generates potential bounding boxes, a convolutional network extracts features, an SVM scores the boxes, a linear model adjusts the bounding boxes, and non-max suppression eliminates duplicate detections. Each stage of this complex pipeline must be precisely tuned independently. Moreover, the resulting system is so ineffective that it takes more than 40 seconds per image at test time [8].

YOLO shares several similarities with R-CNN. Each grid cell proposes potential bounding boxes and scores those boxes using convolutional features. Our system proposes far fewer bounding boxes, only 98 per image compared to about 2000 from Selective Search.
Finally, our system combines these individual components into a single, jointly optimized model.

Conclusion

This paper gives us a review of the YOLO version. We draw the following conclusions.
First, the YOLO version is a lot different. However, they still have certain features in common. Hence, they are still similar.
Second. The YOLO is newfangled algorithm with a simpler structure and shorter training time than R-CNN. There is still a room to ameliorate.

References

[1] Synectic. (March 2015). CCTV in the UK: What we can learn from public attitudes towards surveillance.
[2] Sultana, F., Sufian, A., & Dutta, P. (2020). A review of object detection models based on convolutional neural network. Intelligent Computing: Image Processing Based Applications, 1-16.
[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A.. (2016). You only look once: unified, real-time object detection.
[4] Jamtsho, Y., Riyamongkol, P., & Waranusast, R.. (2019). Real-time bhutanese license plate localization using yolo. ICT Express, 6(2).
[5] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.
[6] Jiang, P. Y., Daji, E., Liu, F. Y., Cai, Y. & Ma, B.. (2022) A Review of Yolo Algorithm Developments. Procedia Computer Science, Volume 199, 2022, (1066-1073).
[7] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. International journal of computer vision, 104(2):154–171, 2013.4.
[8] R. B. Girshick. Fast R-CNN. CoRR, abs/1504.08083, 2015.2, 5, 6, 7.