ISSN No:-2456-2165

Survey on Different Object Detection and Segmentation Methods

Sanket Chandrakant Patel (LDCE ME)
Prof. Pinal Salot (LDCE ME)
L.D. Institute of Engineering

Abstract:- Object Detection is a common problem associated closely with the Computer Vision problem which deals with identifying objects and locating exact positions of certain classes in the image. Interpreting the object positions and localization of classes can be done in various different ways, the most common ones are creating a bounding box around the object and another is marking every pixel in the image which contains the object which is called image segmentation. This survey is based on comparing the two broad classes of object detection algorithms that differ from each other in aspects of a number of stages or steps involved in performing the detection, single-stage detection(YOLO), and two-stage detection (or multi-stage detection) (CNN's). We will also be looking at SOLO architecture which puts light on a completely different approach for segmentation.

Keywords:- *Object Detection, Localization, Detection, Segmentation.*

I. INTRODUCTION

In this paper, we are going to discuss various methods and techniques explored over years by scholars exploring the field of Computer Vision and Pattern Recognition (CVPR) since the 1980s. But in recent years this field has gained an increasing interest of enthusiasts as there is a lot more growth in the related field of automated cars and other sectors both academically as well as real-world applications that are using computer vision to enhance the technology. Also, advances in hardware and software technologies have given rise to the probability of achieving inhumane results in the field. Now, Object Detection and Instance segmentation are not soo easy to solve problem statements as it requires both accuracy and speed. With only any one of these, we cannot truly achieve solutions to real-world problems.

CNN Family

A CNN (Convolutional Neural Network) is a class of deep neural networks that extends the ANN's (Artificial Neural Networks) mostly applied to analyzing visual imagery. ANN tries to simplify and mimics the neural network system of the human brain. As a result, ANN has some fundamental features which are similar to the brain. However, CNN was developed as an enhancement to ANN to satisfy requirements in the image processing area where ANN's cannot give accurate results.

R-CNN

In 2014, the R-CNN(Region-based Convolutional Neural Network) algorithm was proposed as a simple and scalable object detection algorithm with improvements of mean average precision (mAP) by more than 30%. The results were measured on the canonical PASCAL VOC dataset (VOC 2012) achieving an mAP of 53.3%. The R-CNN approach combines two main aspects i.e applying high-capacity convolutional neural networks (CNNs) to bottom-up region proposals for localizing and segments the objects and when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost.

Advantages :

- Using proper CNN Networks one can extract image features automatically.
- The regression model is used to reduce errors of positioning and correct the boundary box predictions.
- Lower Error Rate than Conventional CNN's.

Disadvantages :

- Poor memory management. If there are a lot of features to be extracted then the system can run out of memory.
- More Complicated than most other existing algorithms at the time.
- Training the model can be a lengthy process.
- Slow running speed soo cannot be used for real-time applications.

Fast R-CNN

In 2015, the Fast R-CNN (Region-based Convolutional Neural Network) algorithm was proposed as an improved version of the R-CNN algorithm which aims to improve the image processing efficiency of the R-CNN algorithm. In Comparison, Fast R-CNN trains the very deep VGG16 network 9x faster than R-CNN, and is 213x faster at test-time than R-CNN, and achieves a higher mAP on PASCAL VOC 2012 Data Set. FastR-CNN consists of processes that execute in a pipeline. Firstly, a pre-trained CNN will be applied to both, the classification task and choosing the searching area. Followed by exchanging the max-pooling layer with the ROI(Region of Interest) pooling layer. Finally, Fast R-CNN generates an output which is a discrete probability function for every ROI along with the predicted bounding-box regression model(relative to ROI's).

Advantages :
- Replaces the time-consuming training process of the stage by stage execution with multitasking training.
- Uses ROI pooling to fulfill multi-scale requests.

Disadvantages :
- Time-Consuming algorithm as the core selective search algorithm is slow itself.

Faster R-CNN

In 2016, the Faster R-CNN (Region-based Convolutional Neural Network) algorithm was proposed as an improved version of the Fast R-CNN algorithm replacing the main bottleneck selective search algorithm by region proposal network (RPN's) which uses pre-trained image classification. The region proposal network generates the anchors from the feature maps that are passed to other classification (for finding the area of interest) and regression layers (for predicting bounding boxes).

Advantages :
- First CNN Family algorithm to produce decent real-time results.
- Better accuracy and mAP on outputs.

Disadvantages :
- Reshaping the predicted region proposals before predicting the actual offsets for the bounding box is overhead and the object proposal requires a lot of time.

Mask R-CNN

In 2018, the Mask R-CNN (Region-based Convolutional Neural Network) algorithm was proposed as a conceptually simple, flexible, and general framework for object instance segmentation adding the pixel level masks with pixel image segmentation. It adds an extra parallel layer for generating the masks on each ROIAlign. Another change from Faster R-CNN is in the ROI pooling as it required more accurate bounding boxes for segmentation.

Advantages :
- All the operations i.e Classification, box regressions and mask generations work in parallel which is trending and yields more efficient results.

Disadvantages :
- Needs high computational power as layers work in parallel.

YOLO Family

YOLO is an abbreviation for You only look once. Most of the detection algorithms have to go through the image more than one time to detect the objects in the image and mask the pixels. But YOLO does not need to go through the image multiple times, it detects all the objects in a single scan of the image.

YOLOv1

In the YOLOv1 version, the image is broken down into smaller grids of say TxT, and if the central point of any object is detected then that particular grid is the deciding grid for the object detection. The main idea here is to predict the tensors with the help of CNN Network.

Advantages :
- Faster than other pre-existing solutions.
- Comparatively less false positives are detected in images with complex backgrounds.

Disadvantages :
- If a grid cell contains more than one object, the model will not be able to detect all of them; this is the problem of close object detection that YOLO suffers from.
- The number of objects that can be detected is equal to the number of grids i.e in a 5x5 grid maximum of 25 objects can be detected.

YOLOv2

YOLOv2 is an improved version of YOLOv1. YOLOv2 mainly focused on solving two major drawbacks of YOLOv1, reducing localization errors and improving low recall. YOLOv2 was improved by introducing Batch Normalization, Anchor boxes, and high-resolution classifiers to YOLOv1.

Advantages :
- Improved Speed and accuracy.
- YOLOv2 uses images to train classifiers and also uses images for object detection. This makes it easier to train the detector and improve mAP.

Disadvantages :
- Training the model is lengthy as first, we need to train the classifier network then replace the connected layers with convolutional layers and retrain the model.

YOLOv3

YOLOv3 uses logistic regression to predict the objectiveness score called confidence for each bounding box. YOLOv3 tends to find the highest IOU (Intersection Over Union) for finding the ground truth bounding box for the object. It also differs in the calculations of classification losses. It uses binary cross-entropy loss rather than mean square error.

Advantages :
- Three times faster than previous methods.
- Better detection of smaller objects in the images.
- Detects outputs at three different stages rather than at the final output stage.

Disadvantages :
- Higher Positioning errors due to which low AP scores.

YOLOv4

YOLOv4 was proposed in 2020 with addition or enhancement of features like Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections (CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT) and Mish-activation which are universal features and do not require a specific data set or setup. The paper proposed an efficient and powerful object detection model which is easy to train. The model verifies

the influence of state-of-the-art Bag-of Freebies and Bag-of-Specials methods of object detection during the detector training. Due to which high speed detection is achieved with 43.5 % AP state-of-art results over MS COCO dataset and 65 FPS real-time speed on Tesla V100.

Advantages :
- Modified state-of-the-art methods make them more efficient and suitable for single GPU training.
- High speed detection at 65 FPS on real-time systems.

Disadvantages :
- Introduction of so many layers increases the overhead of deciding the compromise trade off between the mAP and the training and inference speed as to allow the model to run on embedded systems.

YOLOv5

YOLOv5 was proposed later in 2020 with exceptional efficiency improvements but as the matter of controversies for the name YOLOv5 no official paper was published. There are several versions of YOLOv5 itself each providing different features and enhancements on different parameters. YOLOv5 was implemented with pyTorch aiding in achieving great results.

Advantages :
- It is built on top of PyTorch that makes the training & inference process very easy and fast and helps achieve better results.

SOLO Family

SOLO (Segmenting Objects by Locations) is a simple approach for instance segmentation proposed in 2020. For predicting the masks most of the existing methods follow either of two approaches , the “detect-then-segment” strategy as used in Mask R-CNN or predict embedding vectors first then use clustering techniques to group pixels into individual instances. SOLO proposed a completely new approach introducing the concept of “instance categories”, that assigns categories to every pixel within an instance according to the instance’s location and size, thus converting instance segmentation into a single-shot classification-solvable problem.

SOLO

SOLO aims to eliminate the overhead of detecting the bounding boxes for the objects to generate the instance masks. All the existing models either heavily rely on accurate bounding box detection or depend on per-pixel embedding learning and the grouping processing. In contrast, SOLO aims to directly segment instance masks, under the supervision of full instance mask annotations instead of masks in boxes or additional pixel pairwise relations. SOLO uses the concept of Locations where the image is divided into TxT grids leading T^2 central location classes. It also uses Feature Pyramid Networks (FPN) to differentiate objects of different sizes.

Advantages :
- Eliminate the need to predict the bounding boxes and directly generate instance masks.
- SOLO converts the instance segmentation problem into a position aware classification task which is easy to solve.

Disadvantages :
- The results are heavily dependent on optimal selection of the grid sizes.
- Mask predictions affected with the resolution of inputs.

SOLOv2

SOLOv2 was proposed later in 2020 which is based on the SOLO architecture. It proposed several improvements over the previous model. SOLOv2 uses novel matrix non-maximum suppression (NMS) technique which reduces the overhead in inferences significantly. SOLOv2 eliminates three bottlenecks of previous work, inefficient mask representation and learning ,not high enough resolution for finer mask predictions and slow mask NMS(non-maximum suppression). It achieves 38.8% mask AP at 18 fps over MS COCO Dataset.

Advantages :
- Improved inference process with replacing only subset S^2 kernels from the output tensor M rather than directly computing M in memory which is computationally inefficient.
- SOLOv2 is able to predict fine and detailed masks, especially at object boundaries.

Disadvantages :
- The prototype masks exhibit more complex patterns than that in SOLO.

II. CONCLUSION

In the above work, there is a discussion about CNN Family, YOLO Family and emerging SOLO Family. The CNN Family solved the detection and segmentation problem in two steps i.e , detecting object categories (Classification Problem) and then detecting the actual object locations (bounding accurate boxes). The YOLO Family models aims to translate the detection problem into a regression problem and aims to solve the problem in a single frame scan. The SOLO Family architectures are totally box-free providing direct instance segmentation providing state-of-art results with comparatively less computation.

REFERENCES

[1]. Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.
[2]. Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39.6 (2016): 1137-1149.
[3]. He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[4]. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. pp. 1-10
[5]. Redmon, J., Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. pp. 1-23.
[6]. Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
[7]. Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv preprint arXiv:2004.10934 (2020).
[8]. Wang, Xinlong, et al. "SOLOv2: Dynamic and fast instance segmentation." Advances in Neural Information Processing Systems 33 (2020).
[9]. Wang, Xinlong, et al. "Solo: Segmenting objects by locations." European Conference on Computer Vision. Springer, Cham, 2020.