Notes - You Only Look Once (YOLO)

Monday, December 24, 2018 6:24 PM

What is YOLO?
A new approach to object detection. It treats object detection as a regression problem to spatially separated bounding boxes and associated class probabilities.

What are the previous approaches before YOLO?
The approaches used previously have made use of the classifiers to perform detection.

What would the structure of YOLO would look like?
A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. This network is optimized end-to-end directly on detection performance. The bigger YOLO can process images at the rate of 45 frames per second. The other network which is Fast YOLO can achieve a speed of 155 frames per second with the mAP more than double times the previous approaches. YOLO reframes the object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. This algorithm looks at the image only once and predict the objects classes and locations.

What is the analogy between the YOLO and the real life functioning?
Humans glance at an image once and instantly know what objects are in the image, where they are, and how they interact. The YOLO algorithm also tries to implement the same.

What are the significant previous approaches?

1. Deformable Parts Models (DPM)
2. R-CNN

What is Deformable Parts Models (DPM) ?
It is an algorithm that is used before YOLO and it uses the sliding window approach where the classifier is run at evenly spaced locations over the entire image.

What is R-CNN?
They use region proposal methods which first generate the potential bounding boxes in an image and then run a classifier on those proposed boxes. After the classification, post-processing is used to refine the bounding box, eliminate duplicate detections and rescore the box based on other objects in the scene.

What are the advantages of YOLO?
Fast and accurate:
YOLO doesn't have a complex pipeline. With no batch processing, base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps. The video can be processed in real-time with less than 25 milliseconds of latency. YOLO can also achieve more than twice the mean average precision of other real-time systems.

Global inference:
Unlike sliding window and region proposal methods, YOLO reasons globally about the image when making predictions. YOLO sees an entire image during the training and test time so it encodes contextual information about the classes as well their appearance. Fast R-CNN, a top detection method, mistakes background patches in an image for object because it is not able to see the larger context.

Generalizable representations:
After training on natural images, YOLO can also work on the art-work and out performs detection methods like DPM and R-CNN by a wide margin.