YOLO V2

YOLO is an end-to-end real-time target detection system based on deep learning methods. The author has adopted a series of methods to optimize the model structure of YOLO, produced YOLOv2, and reached the state of the art quickly and accurately.

The problem with YOLO's current best target detection system is not enough accuracy. Error item analysis shows that YOLO has a higher error rate in target positioning than Fast R-CNN. Therefore, improvements to YOLO focus on improving positioning accuracy while maintaining classification accuracy.

The authors propose a mechanism for joint training on classification datasets and detection data sets. Use the image of the inspection data set to learn to detect relevant information, such as the Bounding Box coordinate prediction, whether it contains objects and the probabilities belonging to each object. Use class-label-only classification dataset images to expand the types that can be detected.

Monitoring data and classification data are mixed together during training. When the network encounters a picture that belongs to the detection data set, it reversely propagates based on the total loss function of YOLOv2 (including the classification section and the detection section). When the network encounters a picture that belongs to the classification data set, it only back-propagates based on the loss function of the classification part.

This method has some difficulties that need to be solved. The detection dataset only has common objects and abstract labels (not specific), such as "dogs" and "boats." Categorical datasets have a wide and deep range of tags (eg ImageNet has over a hundred dog breeds, including "Norfolk terrier", "Yorkshire terrier", "Bedlington terrier", etc.). Two types of tags must be integrated in a consistent way.

At the same time, YOLOv2 can adapt to different input sizes, adjust detection accuracy and detection speed as needed (it is worth referring). The author integrated the ImageNet data set and the COCO data set, and trained them in a joint training mode, allowing the system to recognize more than 9,000 items. In addition, the author's method of integrating multiple datasets with WordTree can be applied to other computer tasks.