Consistency-based Semi-supervised Learning for Object Detection

Abraham Jose

1 Summary

The author proposes a semi-supervised learning scheme for partially labeled bounding boxes(semi-supervised learning). The CSD(consistency based Semi-supervised learning method for object detection) makes use of consistency and consistency loss between original and flipped images(RPN response in case of the two-stage detector) alongside with supervised training loss to train the model. They also proposes Background Elimination for discarding regions that are in the background which may affect the training process. They extended Consistency Regularization from conventional semi-supervised classification problem to deep learning. The proposed loss ensures consistent classification of the bounding box and the regression of its location.

2 Good points

The strongest point of this paper is that the authors were able to bring the techniques used in conventional computer vision and integrate it to achieve really good semi-supervised results in deep learning and is considered as the class of approaches that yielded ground breaking results in semi-supervised learning. The author has shown great attention to details and deployed the model with Background Elimination to reduce the detection noise from the background that affects the actual labels.

3 Weak points

Proposed background elimination does not work as expected and the labels with similar visual features(like dog and horse) can deter the performance of the model. Consistency treatment of single-stage and two-stage detector are not same as in the former method, we use horizontally flipped images and in the latter one, we uses horizontally flipped response from the region proposals generated by RPN. Performance degradation, when used with fully labeled data, should be justified and the effect of the consistency loss and its regularization effect should be studied.

4 Questions

Why do the consistency model is different for single-stage and two-stage detector. Based on the theory they proposed, should there be a better way to get detection/response for horizontally flipped images in the two-stage detectors?

5 Ideas

To reduce the confusion across objects with similar visual features, we can use a fine graded classifier backbone to the network so that the model will be able to distinct between the fine features. For example using a bi-linear CNN as backbone to extract the features from the images which is used for training.