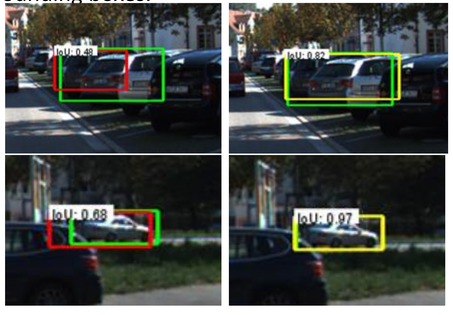
Accurate Single Stage Detector Using Recurrent Rolling Convolution

Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai , Li Xu
SenseTime Group Limited
{rensijie, chenxiaohao, liujianbo, sunwenxiu, pangjiahao, yanqiong, yuwing , xuli}@sensetime.com

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Problems of existing methods

The single stage detectors are usually easier to train and more computationally efficient in production, but failed to generate high quality bounding boxes.



Left column: Previous single stage detector fails to generate bounding boxes of high IoU to the groundtruth; Right column: With the proposed RRC.

Contribution

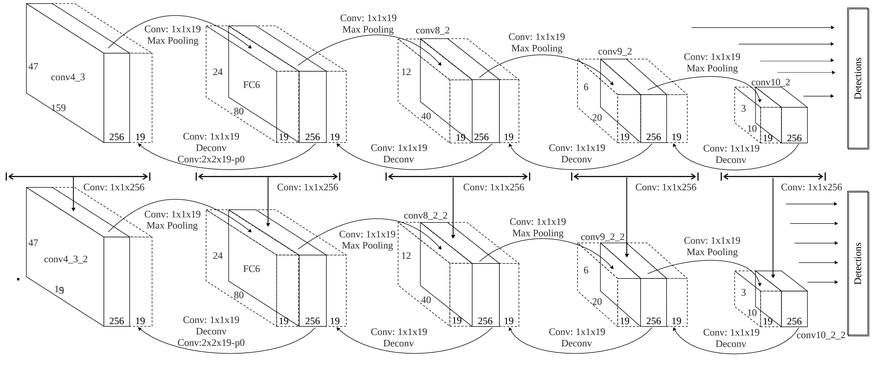
- First, we showed that it is possible to train a single stage detector in the end-to-end fashion to produce very accurate detection results for tasks requiring high localization quality.
- Second, we discovered that the key for improving single stage detector is to recurrently introduce context to the bounding box regression. This procedure can be efficiently implemented with the proposed Recurrent Rolling Convolution architecture.

Overview

Most of the recent successful methods in accurate object detection and localization use some variant of R-CNN style two stage CNN. The single stage detection methods have not been as competitive when evaluated in benchmarks consider mAP for high IoU thresholds. In this paper, we proposed a novel single stage end-to-end trainable object detection network to overcome this limitation. We achieved this by introducing Recurrent Rolling Convolution (RRC) architecture over multi-scale feature maps to construct object classifiers and bounding box regressors which are “deep in context”.

Recurrent Rolling Convolution Architecture

We proposed a novel Recurrent Rolling Convolution (RRC) architecture to improve the localization accuracy in a single stage network. We achieved this by introducing over multi-scale feature maps to construct object classifiers and bounding box regressors which are “deep in context”.



All the features maps (solid boxes) in the first stage were previously computed by the backbone reduced VGG16 network. In each stage, the arrows illustrates the top-down/bottom-up feature aggregation. All the weights of such feature aggregation are shared across stages.

Our Results

Results on the KITTI Car testing set (moderate)

Methods	Car Moderate
SubCNN [20]	89.04%
MS-CNN [2]	89.02%
SDP+RPN [22]	88.85%
Mono3D [3]	88.66%
3DOP [4]	88.64%
RRC(single)	89.85%
RRC(ensemble)	90.19%

Results on the KITTI Car testing set (hard)

Methods	Car Hard
DuEye (anonymous)	86.18%
Genome (anonymous)	85.82%
eagle (anonymous)	85.66%
RV-CNN (anonymous)	85.43%
RRC (ours)	86.97%

Results on the KITTI Pedestrian testing set (moderate)

Methods	Pedestrian Moderate
SubCNN [20]	73.70%
MS-CNN [2]	71.33%
SDP+RPN [22]	70.16%
RRC (ours)	75.33%

Results on the KITTI Cyclist testing set (moderate)

Methods	Cyclist Moderate
SubCNN [20]	71.06%
MS-CNN [2]	75.46%
SDP+RPN [22]	73.74%
RRC (ours)	76.47%

Our code is publicly available