Grid R-CNN

su

2018/1207

Main Contributions

- propose a novel localization framework called Grid R-CNN which substitute traditional regression network by fully convolutional network that preserves spatial information efficiently.

- We design a multi-point supervision form that predicts points in grid to reduce the impact of some inaccurate points. We further propose a feature map level information fusion mechanism that enables the spatially related grid points to obtain incorporated features so thattheir locations can be well calibrated

- We perform extensive experiments and prove that Grid R-CNN framework is widely applicable across different detection frameworks and network architectures with consistent gains. The Grid R-CNN performs even better in more strict localization criterion (e.g. IoU threshold = 0.75). Thus we are confident that our grid guided localization mechanism is a better alternative for regression based localization methods.

Difference

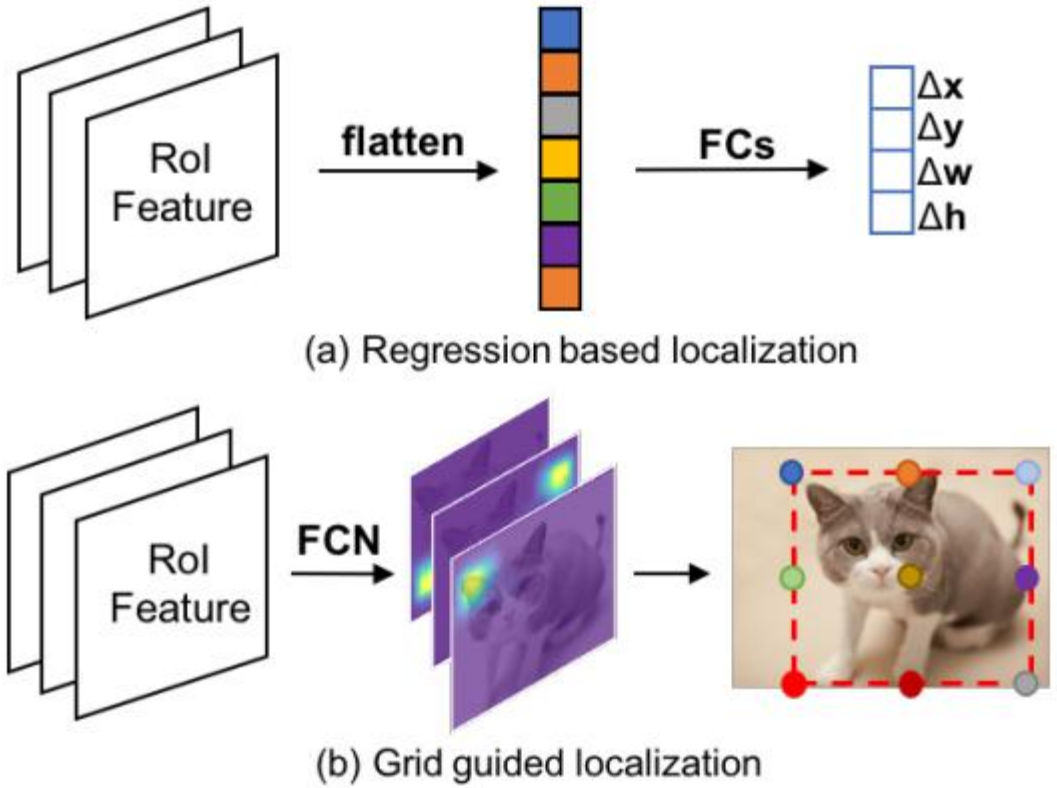


Figure 1. (a) Traditional offset regression based bounding box localization. (b) Our proposed grid guided localization in Grid R-CNN. The bounding box is located by a fully convolutional network.

Since a bounding box has four degrees of freedom, two independent points (e.g. the top left corner and bottom right corner) are enough for localization of a certain object. However the prediction is not easy because the location of the points are not directly corresponding to the local features. For example, the upper right corner point of the cat in Figure 1.b lies outside of the object body and its neighborhood region in the image only contains background,it may share very similar local features with nearby pixels.

To overcome this problem, we design a multi-point supervision formulation. By defining target points in a gird, we have more clues to reduce the impact of inaccurate prediction of some points. For instance, in a typical 3×3 grid points supervision case, the probably inaccurate y-axis coordinate of the top-right point can be calibrated by that of top-middle point which just locates on the boundary of the object. The grid points are effective designs to decrease the overall deviation

Overview of the pipeline of Grid R-CNN

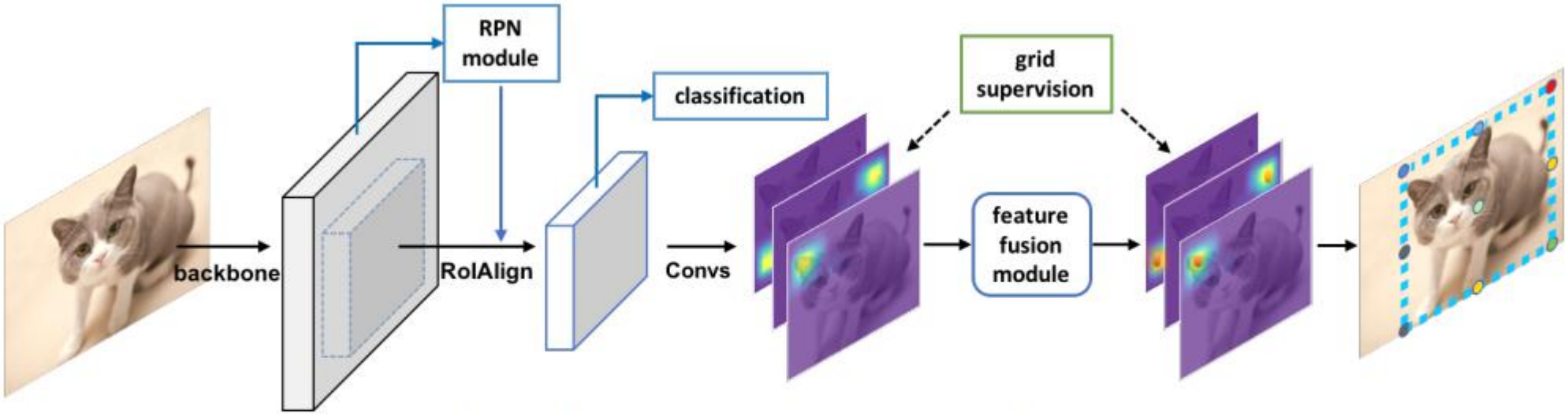


Figure 2. **Overview of the pipeline of Grid R-CNN.** Region proposals are obtained from RPN and used for RoI feature extraction from the output feature maps of a CNN backbone. The RoI features are then used to perform classification and localization. In contrast to previous works with a box offset regression branch, we adopt a grid guided mechanism for high quality localization. The grid prediction branch adopts a FCN to output a probability heatmap from which we can locate the grid points in the bounding box aligned with the object. With the grid points, we finally determine the accurate object bounding box by a feature map level information fusion approach.

Overview of the pipeline of Grid R-CNN

cation on the original image as the grid point. Formally, a point (H_x, H_y) in heatmap will be mapped to the point (I_x, I_y) in origin image by the following equation:

$$I_x = P_x + \frac{H_x}{w_o} w_p$$
$$I_y = P_y + \frac{H_y}{h_o} h_p \quad (1)$$

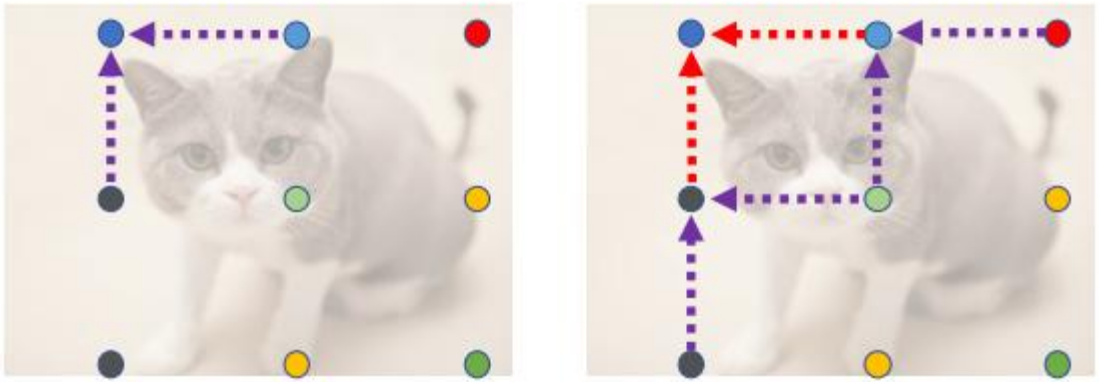
where (P_x, P_y) is the position of upper left corner of the proposal in input image, w_p and h_p are width and height of proposal, w_o and h_o are width and height of output heatmap.

Then we determine the four boundaries of the box of object with the predicted grid points. Specifically, we denote the four boundary coordinates as $B = (x_l, y_u, x_r, y_b)$ representing the left, upper, right and bottom edge respectively. Let g_j represent the j-th grid point with coordinate (x_j, y_j) and predicted probability p_j,. Then we define E_i as the set of indices of grid points that are located on the i-th edge, i.e., $j \in E_i$ if g_j lies on the i-th edge of the bounding box. We have the following equation to calculate B with the set of g:

$$x_l = \frac{1}{N} \sum_{j \in E_1} x_j p_j, \qquad y_u = \frac{1}{N} \sum_{j \in E_2} y_j p_j$$
$$x_r = \frac{1}{N} \sum_{j \in E_3} x_j p_j, \qquad y_b = \frac{1}{N} \sum_{j \in E_4} y_j p_j \quad (2)$$

Taking the upper boundary y_u as an example, it is the probability weighted average of y axis coordinates of the three upper grid points.

Grid Points Feature Fusion



(a) (b)

Figure 3. An illustration of the 3×3 case of grid points feature fusion mechanism acting on the top left grid point. The arrows represent the spatial information transfer direction. (a) First order feature fusion, feature of the point can be enhanced by fusing features from its adjacent points. (b) The second order feature fusion design in Grid R-CNN.

$$F_i' = F_i + \sum_{j \in S_i} T_{j \to i}(F_j) \tag{3}$$

Based on F_i' for each grid point, a second order of fusion is then performed with new conv layers $T^+_{j \to i}$ that don't share parameters with those in first order of fusion. And the second order fused feature map F_i'' is utilized to output the final heatmap for the grid point location prediction. The second order fusion enables an information transfer in the range of 2 (L_1 distance). Taking the upper left grid point in 3×3 grids as an example (shown in Figure 3.b), it synthesizes the information from five other grid points for reliable calibration.

Extended Region Mapping

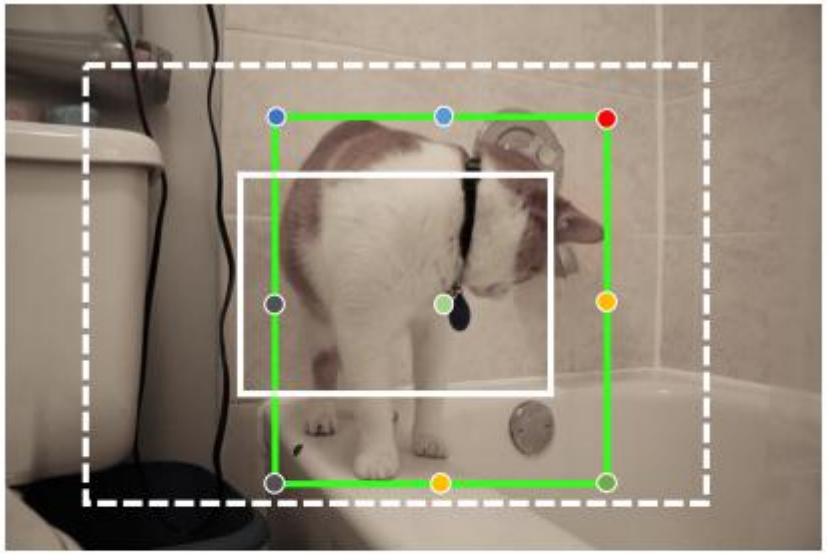


Figure 4. Illustration of the extended region mapping strategy. The small white box is the original region of the RoI and we extend the representation region of the feature map to the dashed white box for higher coverage rate of the grid points in the the ground truth box which is in green.

The extended region mapping is formulated as a modification of Equation 1:

$$I_x^{'} = P_x + \frac{4H_x - w_o}{2w_o} w_p$$
$$I_y^{'} = P_y + \frac{4H_y - h_o}{2h_o} h_p \quad (4)$$

After the new mapping, all the target grid points of the positive proposals (which have an overlap larger than 0.5 with ground truth box) will be covered by the corresponding region of the heatmap.

Result

method	AP	$AP_{.5}$	$AP_{.75}$
regression	37.4	59.3	40.3
2 points	38.3	57.3	40.5
4-point grid	38.5	57.5	40.8
9-point grid	38.9	58.2	41.2

Table 1. Comparison of different grid points strategies in Grid R-CNN. Experiments show that more grid points bring performance gains.

method	AP	$AP_{.5}$	$AP_{.75}$
w/o fusion	38.9	58.2	41.2
bi-directional fusion [26]	39.2	58.2	41.8
first order feature fusion	39.2	58.1	41.9
second order feature fusion	39.6	58.3	42.4

Table 2. Comparison of different feature fusion methods. Bi-directional feature fusion, first order feature fusion and second order fusion all demonstrate improvements. Second order fusion achieves the best performance with an improvement of 0.7% on AP.

Result

method	backbone	AP	$AP_{.5}$	$AP_{.75}$	AP_S	AP_M	AP_L
Faster R-CNN	ResNet-50	33.8	55.4	35.9	17.4	37.9	45.3
Grid R-CNN	ResNet-50	**35.9**	54.0	38.0	18.6	40.2	47.8
Faster R-CNN w FPN	ResNet-50	37.4	59.3	40.3	21.8	40.9	47.9
Grid R-CNN w FPN	ResNet-50	**39.6**	58.3	42.4	22.6	43.8	51.5
Faster R-CNN w FPN	ResNet-101	39.5	61.2	43.1	22.7	43.7	50.8
Grid R-CNN w FPN	ResNet-101	**41.3**	60.3	44.4	23.4	45.8	54.1

Table 5. Bounding box detection AP on COCO *minival*. Grid R-CNN outperforms both Faster R-CNN and FPN on ResNet-50 and ResNet-101 backbone.

Result

method	backbone	AP	$AP_{.5}$	$AP_{.75}$	AP_S	AP_M	AP_L
YOLOv2 [14]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD-513 [15]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD-513 [16]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
RefineDet512 [17]	ResNet101	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet800 [18]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
CornerNet	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
Faster R-CNN+++ [8]	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [4]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN w TDM [19]	Inception-ResNet-v2 [22]	36.8	57.7	39.2	16.2	39.8	52.1
D-FCN [20]	Aligned-Inception-ResNet	37.5	58.0	-	19.4	40.1	52.5
Regionlets [21]	ResNet-101	39.3	59.8	-	21.7	43.7	50.9
Mask R-CNN [5]	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2
Grid R-CNN w FPN (ours)	ResNet-101	41.5	60.9	44.5	23.3	44.9	53.1
Grid R-CNN w FPN (ours)	ResNeXt-101	43.2	63.0	46.6	25.1	46.5	55.2

Table 6. Comparison with state-of-the-art detectors on COCO *test-dev*.

Result

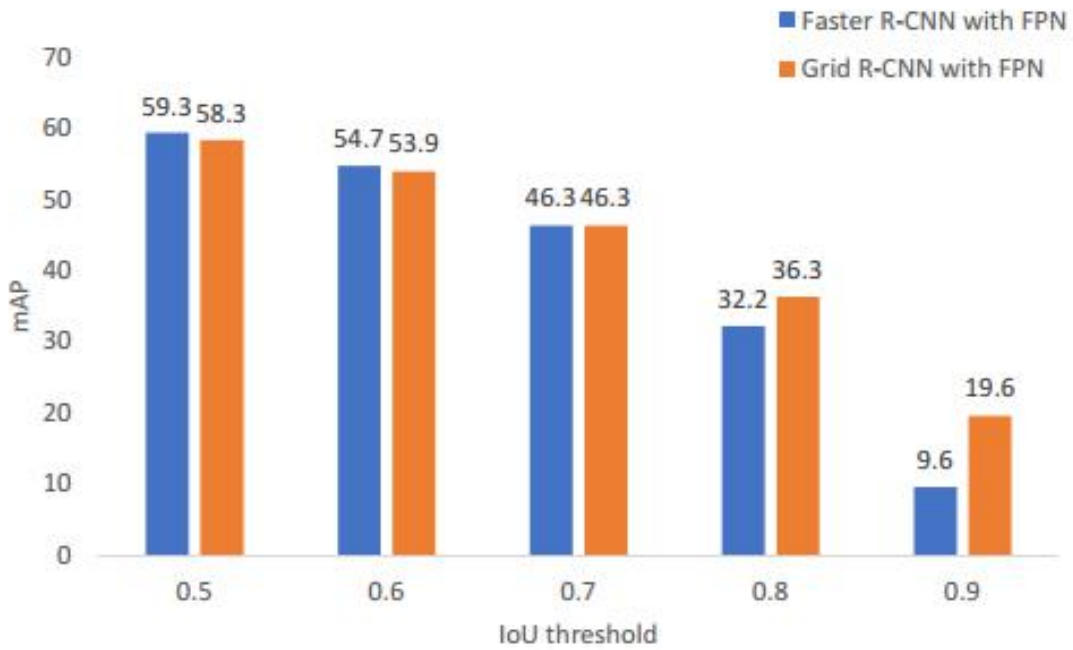


Figure 5. AP results across IoU thresholds from 0.5 to 0.9 with an interval of 0.1.