VarifocalNet

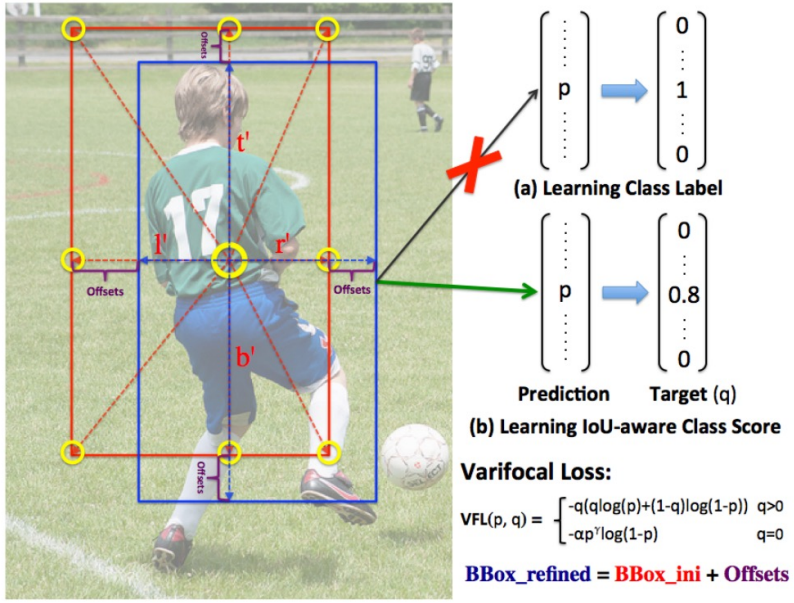
Paper review

Sangho Kim

Table of Contents

Introduction to Ideas

- In this paper, we propose to learn an IoU-aware Classification Score (IACS) as a joint representation of object presence confidence and localization accuracy.
- We design a new loss function, named Varifocal Loss
- Also we propose a new star-shaped bounding box feature representation for IACS prediction and bounding box refinement
- We build an IoU-aware dense object detector based on the FCOS+ATTS architecture, that we call VarifocalNet or VFNet



Introduction to Ideas

- Generally, we classification score is used to rank the bounding box in NMS
- However, this harms the detection performance, because the classification score is not always a good estimate
- To solve the problem, existing dense detectors predict either an additional IoU score or a centerness score as the localization accuracy estimation
- These methods can alleviate the misalignment problem between classification score and the localization accuracy
- However, they are sub-optimal because multiplying the two imperfect prediction may lead to a worse rank basis
- To overcome these shortcomings, we would like to ask:
    - Instead of predicting an additional localization accuracy score, can we merge it into the classification score?

Table of Contents

Related Works

- Focal Loss

$$\mathbf{FL}(p) = -(1-p_t)^{\gamma}\log(p_t), p_t = \begin{cases} p, & \text{when } y = 1 \\ 1-p, & \text{when } y = 0 \end{cases}$$

- Where $y \in \{1, 0\}$ specifies the ground-truth class and $p \in [0,1]$ denotes the estimated probability for the class
- FL consists of a standard cross entropy part $-\log(p_t)$ and a dynamically scaling factor part $(1-p_t)^{\gamma}$
- The scaling factor automatically down-weights the contribution of easy examples during training and rapidly focuses the model on hard examples

Related Works

- Generalized Focal Loss
- Qaulity Focal Loss part in GFL

$$\mathbf{QFL}(\sigma) = -\left|y-\sigma\right|^{\beta}\left((1-y)\log(1-\sigma)+y\log(\sigma)\right).$$

- The scaling factor part $(1-p_t)^{\gamma}$ is generalized into the absolute distance between the estimation σ and its continuous label y, i.e., $|y-\sigma|^{\beta} \quad (\beta > 0)$
- Similar to FL, the scaling factor term of QFL behaves as a modulating factor
- When the quality estimation of example is inaccurate, this term is relatively large, thus it pays more attention to learning hard example
- As the quality estimation becomes accurate, i.e., $\sigma \rightarrow y$, the factor goes to 0 and the loss for well-estimated examples is down-weighted, in which the parameter β controls the down-weighting rate

Related Works

- The difference from Generalized Focal Loss
- We emphasize first that our varifocal loss is a distinct function from the GFL
- It weights positive and negative examples asymmetrically, whereas the GFL deals with them equally
- And experiment results show that our varifocal loss performs better than the GFL

Table of Contents

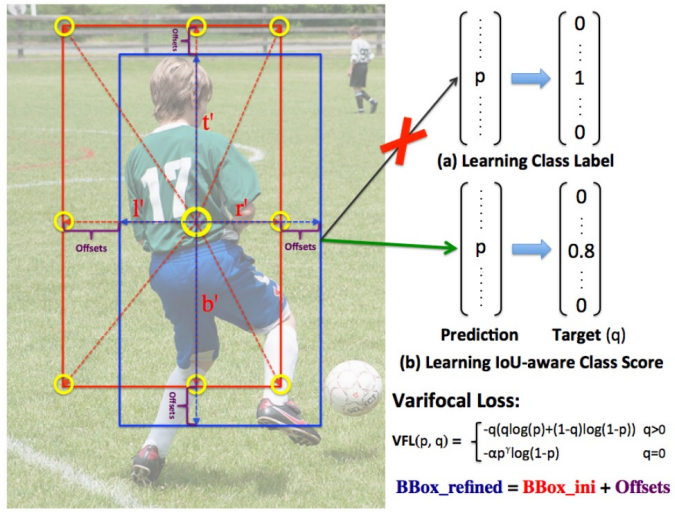
VarifocalNet

- We build a new dense detector VFNet based on the FCOS + ATTS with the centerness branch removed
- We define the IACS as a scalar element of a classification score vector, in which the value at the ground-truth class label position is the IoU between the predicted bounding box and its ground truth, and 0 at other positions
- We define Varifocal Loss as:

$$\mathrm{VFL}(\mathrm{p},\mathrm{q}) = \begin{cases} -\mathrm{q}(\mathrm{q}\log(\mathrm{p}) + (1-\mathrm{q})\log(1-\mathrm{p})) \\ -\alpha \mathrm{p}^{\gamma}\log(1-\mathrm{p}) \end{cases}$$

- Where p is the predicted IACS and q is the target score
- For a foreground point, q for its ground-truth class is set as the IoU between the predicted box and its ground-truth and 0 otherwise, whereas for a background point, the target q for all classes is 0
- The varifocal loss only reduces the loss contribution from negative examples ($q = 0$) by scaling their losses with a factor p^{γ} and does not down-weight positive examples ($q > 0$) in the same way

VarifocalNet



- Star-shaped Box Feature Representation

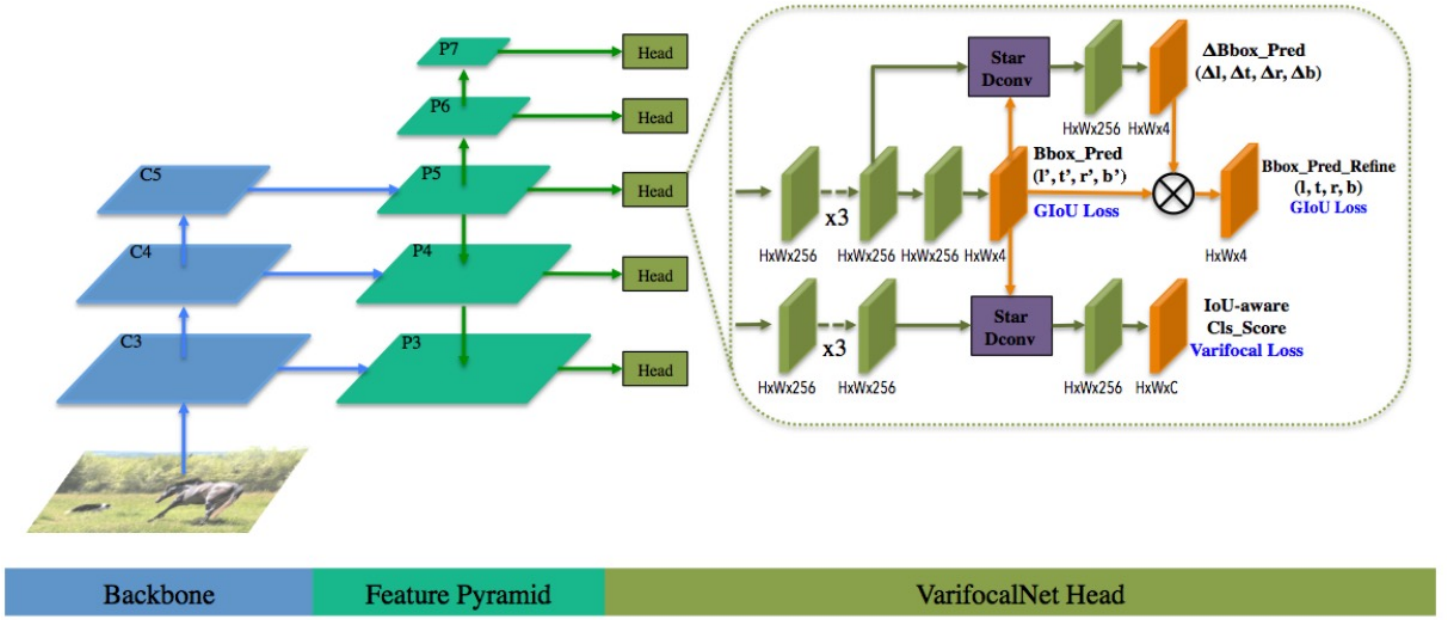
- We design a star-shaped bounding box feature representation for IACS prediction
- It uses the feature at nine fixed sampling points to represent a bounding box with the deformable convolution
- This new representation can capture the geometry of a bounding box and its nearby contextual information
- Given a sampling location (x, y) on the image plane, we first regress an initial bounding box from it with 3×3 convolution
- Following the FCOS, this bounding box is encoded by a $4D$ vector (l', t', r', b') which means the distance from the location (x, y) to the left, top, right and bottom side of the bounding box respectively
- Since these points are manually selected without additional prediction burden, our new representation is computation efficient

VarifocalNet

- Bounding Box Refinement

- The existing technique in object detection, it is not widely adopted in dense detector due to the lack of an efficient
- With our new star representation, we can now adopt it in dense detectors without losing efficiency
- We model the bounding box refinement as a residual learning problem
- For an initially regressed bounding box (l', t', r', b'), we first extract the star-shaped representation to encode it
- Then, based on the representation, we learn four distance scaling factors $(\Delta l, \Delta t, \Delta r, \Delta b)$ to scale the initial distance vector, so that the refined bounding box that is represented by $(\mathrm{l}, \mathrm{t}, r, b) = (\Delta l \times l', \Delta \mathrm{t} \times t', \Delta r \times r', \Delta b \times b')$ is closer to the ground truth

VarifocalNet



- Varifocal Net

- The VFNet head consists of two subnetworks
- The localization subnet performs bounding box regression and subsequent refinement
- It takes as input the feature map from each level of the FPN and first applies three 3×3 conv layers with ReLU activations
- One branch of the localization subnet convolves the feature map again and then outputs a $4D$ distance vector $(\Delta l, \Delta t, \Delta r, \Delta b)$ per spatial location which represents the initial box
- The other branch applies a star-shaped deformable convolution to the nine feature sampling points
- The other subnet aims to predict the IACS that has the similar structure to the localization subnet

VarifocalNet

- Loss Function

$$\begin{aligned}\mathrm{Loss} = & \frac{1}{\mathrm{N_{pos}}} \sum_{\mathrm{i}} \sum_{\mathrm{c}} \mathrm{VFL}(\mathrm{p_{c,i}},\ \mathrm{q_{c,i}}) \\ & + \frac{\lambda_0}{\mathrm{N_{pos}}} \sum_{\mathrm{i}} \mathrm{q_{c^*,i}} \mathrm{L_{bbox}}(\mathrm{bbox'_i},\ \mathrm{bbox^*_i}) \\ & + \frac{\lambda_1}{\mathrm{N_{pos}}} \sum_{\mathrm{i}} \mathrm{q_{c^*,i}} \mathrm{L_{bbox}}(\mathrm{bbox_i},\ \mathrm{bbox^*_i})\end{aligned}$$

- Where $p_{c,i}$ and $q_{c,i}$ denote the predicted and target IACS respectively for the class c at the location i on each feature map of FPN
- L_{bbox} is the GIoU loss, and $bbox'_i$, $bbox_i$ and $bbox^*_i$ represent the initial, refined and ground truth respectively
- λ_0 and λ_1 are the balance weights for L_{bbox} and are empirically set as 1.5 and 2.0 respectively
- N_{pos} is the number of foreground points and is used to normalize the total loss

Table of Contents

Experiments

	FCOS+ATSS									
w/ctr		✓	✓	✓		✓		✓		✓
gt_ctr			✓							
gt_ctr_iou				✓						
gt_bbox					✓	✓				
gt_cls							✓	✓		
gt_cls_iou									✓	✓
AP	38.5	39.2	41.1	43.5	56.1	56.3	43.1	58.1	**74.7**	67.4

Table 1: Performance of the FCOS+ATSS on COCO `val2017` with different oracle predictions. W/ctr means using the centerness score in inference. Please see the text for the meaning of other abbreviations.

Experiments

γ	α	q weighting	AP	AP_{50}	AP_{75}
1.0	0.50	✓	41.2	59.2	44.7
1.5	0.75	✓	41.5	59.7	45.1
2.0	0.75	✓	**41.6**	59.5	45.0
2.0	0.75		41.2	59.1	44.4
2.5	1.25	✓	41.5	59.4	45.2
3.0	1.00	✓	41.3	59.0	44.7

Table 2: Peformance of the VFNet when changing the hyper-parameters (α, γ) of the varifocal loss. q weighting means weighting the loss of the positive example with the learning target q.

Experiments

VFL	Star Dconv	BBox Refinement	AP	AP_{50}	AP_{75}
			39.0	57.7	41.8
✓			40.1	58.5	43.4
✓	✓		40.7	59.0	44.0
✓	✓	✓	**41.6**	59.5	45.0
FCOS+ATSS			39.2	57.3	42.4

Table 3: Individual contribution of the components in our method. The first row represents the results of the raw VFNet trained with the focal loss [8].

Experiments

Method	Backbone	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Anchor-based multi-stage:								
Faster R-CNN [3]	X-101		40.3	62.7	44.0	24.4	43.7	49.8
Libra R-CNN [40]	R-101		41.1	62.1	44.7	23.4	43.7	52.5
Mask R-CNN [4]	X-101		41.4	63.4	45.2	24.5	44.9	51.8
R-FCN [41]	R-101		41.4	63.4	45.2	24.5	44.9	51.8
TridentNet [42]	R-101		42.7	63.6	46.5	23.9	46.6	56.6
Cascade R-CNN [17]	R-101		42.8	62.1	46.3	23.7	45.5	55.2
SNIP [43]	R-101		43.4	65.5	48.4	27.2	46.5	54.9
Anchor-based one-stage:								
SSD512 [7]	R-101		31.2	50.4	33.3	10.2	34.5	49.8
YOLOv3 [6]	DarkNet-53		33.0	57.9	34.4	18.3	35.4	41.9
DSSD513 [44]	R-101		33.2	53.3	35.2	13.0	35.4	51.1
RefineDet [35]	R-101		36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet [8]	R-101		39.1	59.1	42.3	21.8	42.7	50.2
FreeAnchor [18]	R-101		43.1	62.2	46.4	24.5	46.1	54.8
GFL [32]	R-101-DCN		47.3	66.3	51.4	28.0	51.1	59.2
GFL [32]	X-101-32x4d-DCN		48.2	67.4	52.6	29.2	51.7	60.2
EfficientDet-D6 [45]	B6	5.3†	51.7	71.2	56.0	34.1	55.2	64.1
EfficientDet-D7 [45]	B6	3.8†	52.2	71.4	56.3	34.8	55.5	64.6
Anchor-free key-point:								
ExtremeNet [22]	Hourglass-104		40.2	55.5	43.2	20.4	43.2	53.1
CornerNet [20]	Hourglass-104		40.5	56.5	43.1	19.4	42.7	53.9
Grid R-CNN [46]	X-101		43.2	63.0	46.6	25.1	46.5	55.2
CenterNet [20]	Hourglass-104		44.9	62.4	48.1	25.6	47.4	57.4
RepPoints [24]	R-101-DCN		45.0	66.1	49.0	26.6	48.6	57.5
Anchor-free one-stage:								
FoveaBox [15]	X-101		42.1	61.9	45.2	24.9	46.8	55.6
FSAF [27]	X-101-64x4d		42.9	63.8	46.3	26.6	46.2	52.7
FCOS [9]	R-101		43.0	61.7	46.3	26.0	46.8	55.0
SAPD [28]	R-101		43.5	63.6	46.5	24.9	46.8	54.6
SAPD [28]	R-101-DCN		46.0	65.9	49.6	26.3	49.2	59.6
Baseline:								
ATSS [12]	R-101	17.5	43.6	62.1	47.4	26.1	47.0	53.6
ATSS [12]	X-101-64x4d	8.9	45.6	64.6	49.7	28.5	48.9	55.6
ATSS [12]	R-101-DCN	13.7	46.3	64.7	50.4	27.7	49.8	58.4
ATSS [12]	X-101-64x4d-DCN	6.9	47.7	66.5	51.9	29.7	50.8	59.4
Ours:								
VFNet	R-50	19.3	44.3/44.8	62.5/63.1	48.1/48.7	26.7/27.2	47.3/48.1	54.3/54.8
VFNet	R-101	15.6	46.0/46.7	64.2/64.9	50.0/50.8	27.5/28.4	49.4/50.2	56.9/57.6
VFNet	X-101-32x4d	13.1	46.7/47.6	65.2/66.1	50.8/51.8	28.3/29.4	50.1/50.9	57.3/58.4
VFNet	X-101-64x4d	9.2	47.4/48.5	65.8/67.0	51.5/52.6	29.5/30.1	50.7/51.7	58.1/59.7
VFNet	R2-101 [47]	13.0	48.4/49.3	66.9/67.6	52.6/53.5	30.3/30.5	52.0/53.1	59.2/60.5
VFNet	R-50-DCN	16.3	47.3/48.0	65.6/66.4	51.4/52.3	28.4/29.0	50.3/51.2	59.4/60.4
VFNet	R-101-DCN	12.6	48.4/49.2	66.7/67.5	52.6/53.7	28.9/29.7	51.7/52.6	61.0/62.4
VFNet	X-101-32x4d-DCN	10.1	49.2/50.0	67.8/68.5	53.6/54.4	30.0/30.4	52.6/53.2	62.1/62.9
VFNet	X-101-64x4d-DCN	6.7	49.9/50.8	68.5/69.3	54.3/55.3	30.7/31.6	53.1/54.2	62.8/64.4
VFNet	R2-101-DCN [47]	10.3	50.4/51.3	68.9/69.7	54.7/55.8	31.2/31.9	53.7/54.7	63.3/64.4
VFNet-X-800	R2-101-DCN [47]	8.0	53.7	71.6	58.7	34.4	57.5	**67.5**
VFNet-X-1200	R2-101-DCN [47]	4.2	**55.1**	**73.0**	**60.1**	**37.4**	**58.2**	67.0

Table 4: Performance (single-model single-scale) comparison with state-of-the-art detectors on MS COCO `test-dev`. VFNet consistently outperforms the strong baseline ATSS by ~2.0 AP. Our best model VFNet-X-1200 reaches 55.1 AP, achieving the new stat-of-the-art. 'R': ResNet. 'X': ResNeXt. 'R2': Res2Net. 'DCN': Deformable convolution network. '/' separates results of the MSTrain image scale range 1333×[640:800] / 1333×[480:960]. FPSs with † are from papers.

Experiments

Method	AP	AP_{50}	AP_{75}
RetinaNet [8] + FL	36.5	55.5	38.8
RetinaNet [8] + GFL	37.3	56.4	40.0
RetinaNet [8] + VFL	**37.4**	56.5	40.2
FoveaBox [15] + FL	36.3	56.3	38.3
FoveaBox [15] + GFL	36.9	56.0	39.7
FoveaBox [15] + VFL	**37.2**	56.2	39.8
RepPoints [24] + FL	38.3	59.2	41.1
RepPoints [24] + GFL	39.2	59.8	42.5
RepPoints [24] + VFL	**39.7**	59.8	43.1
ATSS [12] + FL	39.3	57.5	42.5
ATSS [12] + GFL	39.8	57.7	43.2
ATSS [12] + VFL	**40.2**	58.2	44.0
VFNet + FL	40.0	58.0	43.2
VFNet + GFL	41.1	58.9	42.2
VFNet + VFL	**41.6**	59.5	45.0

Table 5: Comparison of performances when applying the focal loss (FL) [8], the generalized focal loss (GFL) [32] and our varifocal loss (VFL) to existing popular dense object detectors and our VFNet.