Generalized Focal Loss

Paper Review

Sangho Kim

Table of Contents

The limitation of exsiting works

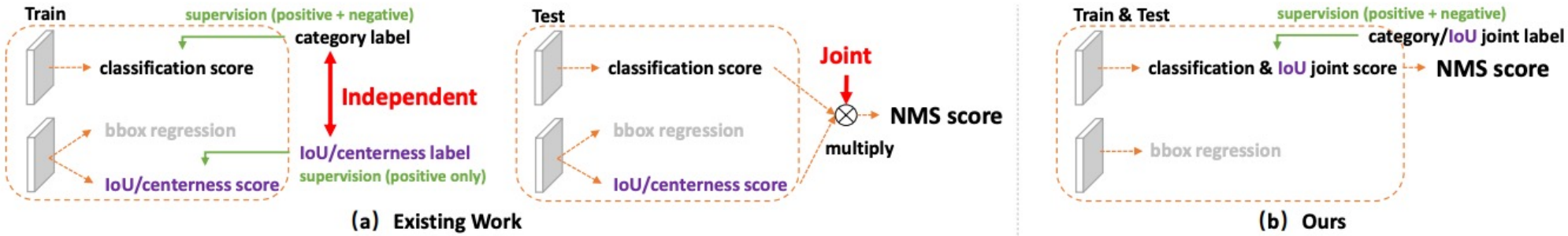
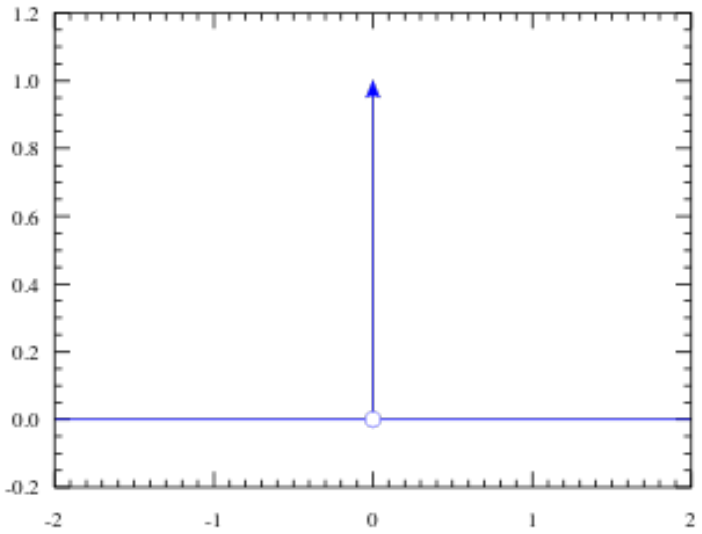
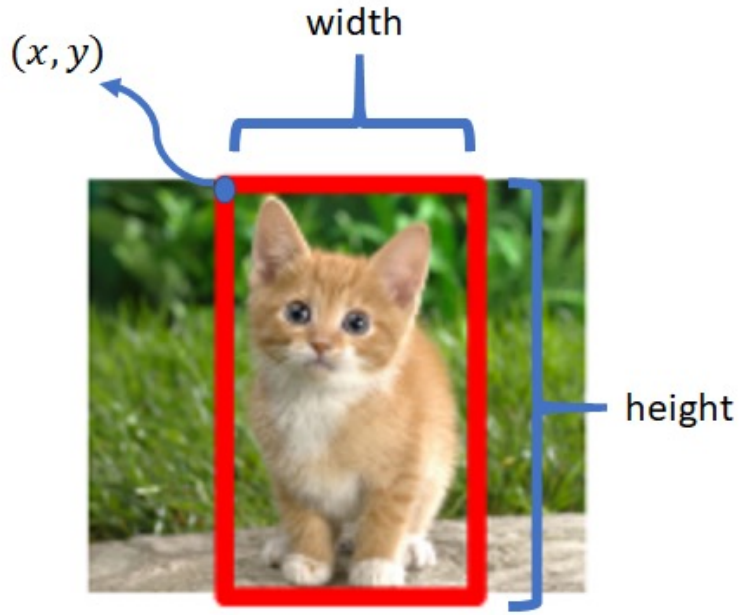


Figure 1: Comparisons between existing separate representation and proposed joint representation of classification and localization quality estimation. (a): Current practices [12, 26, 29, 35, 31] for the separate usage of the quality branch (i.e., IoU or centerness score) during training and test. (b): Our joint representation of classification and localization quality enables high consistency between training and inference.

- In recent dense detectors, the localization estimation and classification score are usually trained independently but compositely utilized during inference.
- These result in a gap between training and test, and would degrade the detection performance.

The limitation of exsiting works





- The widely used bounding box representation can be viewed as Dirac delta distribution.

The limitation of exsiting works

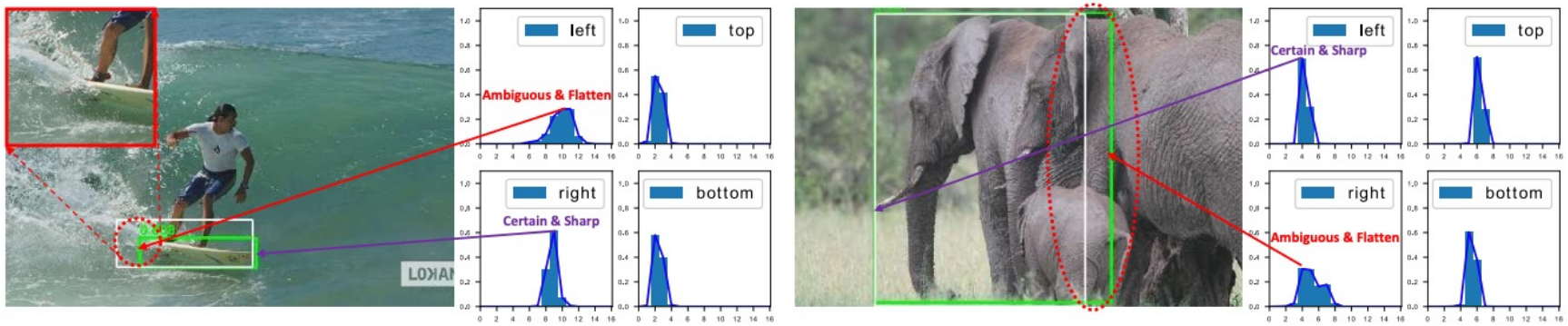
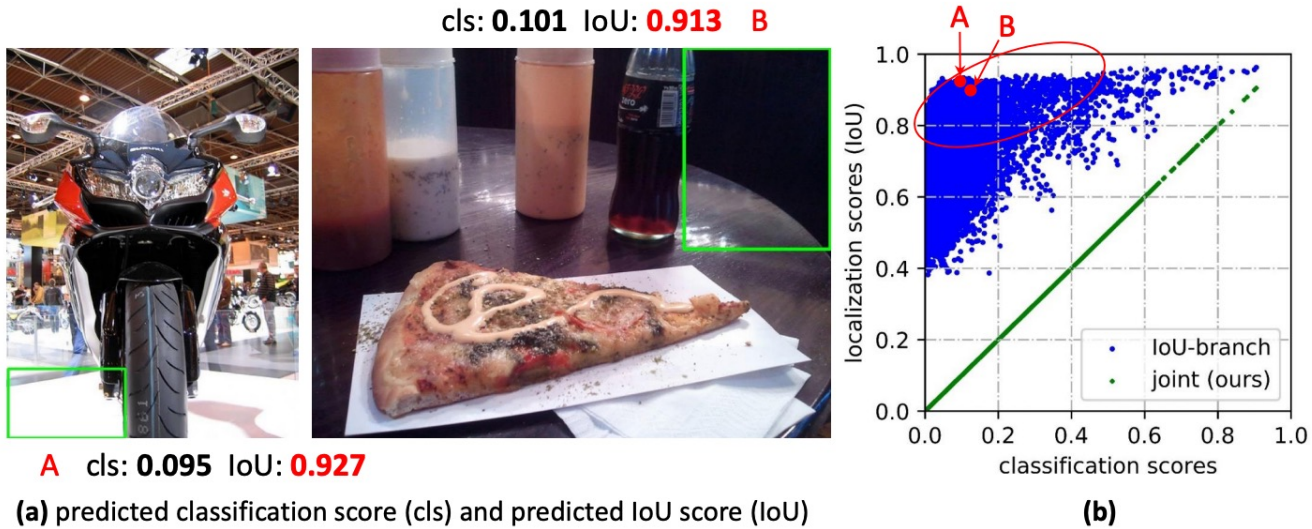


Figure 3: Due to occlusion, shadow, blur, etc., the boundaries of many objects are not clear enough, so that the ground-truth labels (white boxes) are sometimes not credible and Dirac delta distribution is limited to indicate such issues. Instead, the proposed learned representation of General distribution for bounding boxes can reflect the underlying information by its shape, where a flatten distribution denotes the unclear and ambiguous boundaries (see red circles) and a sharp one stands for the clear cases. The predicted boxes by our model are marked green.

- However, it fails to consider the ambiguity and uncertainty in dataset.
- In fact, the real distribution can be more arbitrary and flexible like our works.

The limitation of exsiting works



(a) predicted classification score (cls) and predicted IoU score (IoU) **(b)**

- To address these problems, we design new representations for the bounding boxes and their localization quality.
- We propose to merge it with the classification score in to a single and unified representation.
- We unify classification score and IoU score in to a joint and single variable, which can be trained in an end-to-end fashion.
- As a result, it eliminates the training-test inconsistency and enables the strongest correlation between localization quality and classification.

The limitation of exsiting works

- We propose to represent the General distribution of box locations.
- Consequently, we can obtain more reliable and accurate bounding box estimations, while being aware of a variety of their underlying distributions
- We demonstrate three advantages of GFL
  - 1. It bridges the gap between training and test.
  - 2. It well models the flexible underlying distribution for bounding boxes.
  - 3. The performance of one-stage detectors can be consistently boosted without introducing additional overhead.

Table of Contents

Review Focal Loss

- Focal Loss (FL)

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

- 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.

Table of Contents

Generalized Focal Loss

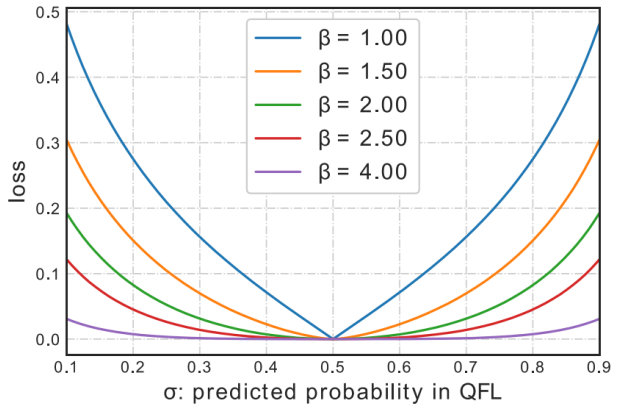
- Quality Focal Loss (QFL)

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

- The cross entropy part $-\log(p_t)$ is expanded into its complete version $-\big((1-y)\ \log(1-\sigma) + y\ log(\sigma)\big)$
- 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., $|y-\sigma|^{\beta}\ \ (\beta > 0)$

Generalized Focal Loss

- Quality Focal Loss (QFL)



- Similar to FL, the term $|y - \sigma|^{\beta}$ of QFL behaves as a modulating factor
- When the quality estimation of an 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., $\sigma \to y$, 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 smoothly

Generalized Focal Loss

- Distribution Focal Loss (DFL)

$$\mathbf{DFL}(\mathcal{S}_i, \mathcal{S}_{i+1}) = -\big((y_{i+1} - y)\log(\mathcal{S}_i) + (y - y_i)\log(\mathcal{S}_{i+1})\big).$$

- Where $S(\cdot)$ is a softmax layer consisting of $n + 1$ units.
- DFL forces the network to rapidly focus on the values near label y, by explicitly enlarging the probabilities of y_i and y_{i+1} (nearest two to y, $y_i \leq y \leq y_{i+1}$).
- As the learning of bounding boxes are only for positive samples without the risk of class imbalance.

Generalized Focal Loss

- Generalized Focal Loss (GFL)

$$\mathcal{L} = \frac{1}{N_{pos}} \sum_{z} \mathcal{L}_{\mathcal{Q}} + \frac{1}{N_{pos}} \sum_{z} \mathbf{1}_{\{c_z^* > 0\}} \left(\lambda_0 \mathcal{L}_{\mathcal{B}} + \lambda_1 \mathcal{L}_{\mathcal{D}} \right)$$

- Where L_Q is QFL and L_D is DFL.
- Typically, L_B denotes the bounding box regression loss.
- N_{pos} stands for the number of positive samples.
- λ_0 (typically 2 as default) and λ_1 (practically $\frac{1}{4}$) are the balance weights for L_Q and L_D, respectively.
- $\mathbf{1}_{\{c_z^* > 0\}}$ is the indicator function, being 1 if $c_z^* > 0$ and 0 otherwise.

Table of Contents

Experiments

Type	FCOS [26]						ATSS [31]					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
w/o quality branch	37.8	56.2	40.8	21.2	42.1	48.2	38.0	56.5	40.7	20.6	42.1	49.1
centerness-branch [26]	38.5	56.8	41.6	22.4	42.4	49.1	39.2	57.4	42.2	23.0	42.8	51.1
IoU-branch [29, 12]	38.7	56.7	**42.0**	21.6	43.0	50.3	39.6	57.6	**43.0**	**23.3**	43.7	51.2
centerness-guided [28]	37.9	56.7	40.7	21.2	42.1	49.4	38.2	56.2	41.0	21.5	41.9	49.7
IoU-guided [28]	38.2	57.0	41.1	**22.5**	42.2	48.9	38.9	57.4	41.8	22.8	42.4	50.6
joint w/ QFL **(ours)**	**39.0**	**57.8**	41.9	22.0	**43.1**	**51.0**	**39.9**	**58.5**	**43.0**	22.4	**43.9**	**52.7**

(a) **Comparisons between separate/implicit and joint representation (ours)**: The joint representation optimized by QFL achieves better performance than other counterparts. We also observe that the quality predictions (especially IoU scores) are necessary for obtaining competitive AP.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FoveaBox [13]	36.4	**55.8**	38.8	19.4	40.4	47.7
FoveaBox [13] + joint w/ QFL	**37.0**	55.7	**39.6**	**20.2**	**41.2**	**48.8**
RetinaNet [18]	35.6	55.5	38.1	20.1	39.4	46.8
RetinaNet [18] + joint w/ QFL	**36.4**	**56.3**	**39.1**	**20.4**	**40.0**	**48.7**
SSD512 [20]	29.4	49.1	30.6	11.4	34.1	44.9
SSD512 [20] + joint w/ QFL	**30.2**	**50.3**	**31.7**	**13.3**	**34.4**	**45.5**

(b) **Applying joint representations with QFL to other one-stage detectors**: About 0.6-0.8 % AP gains are obtained without any additional overhead for inference.

β (QFL)	AP	AP_{50}	AP_{75}
0	37.6	55.4	40.3
1	39.0	58.1	41.7
2	**39.9**	**58.5**	**43.0**
2.5	39.7	58.1	42.7
4	38.2	55.4	41.6

(c) **Varying β for QFL based on ATSS**: $\beta = 2$ performs best.

Table 1: Study on QFL (ResNet-50 backbone). All experiments are reproduced in mmdetection [3] and validated on COCO `minival`.

Experiments

Prior Distribution	FCOS [26]						ATSS [31]					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Dirac delta [26, 31]	38.5	56.8	41.6	22.4	42.4	49.1	39.2	**57.4**	42.2	23.0	42.8	51.1
Gaussian [10, 4]	38.6	56.5	41.6	21.7	42.5	50.0	39.3	57.0	42.4	**23.6**	42.9	51.0
General **(ours)**	38.8	56.6	42.0	22.5	42.9	49.8	39.3	57.1	42.5	23.5	43.0	**51.2**
General w/ DFL **(ours)**	**39.0**	**57.0**	**42.3**	**22.6**	**43.0**	**50.6**	**39.5**	57.3	**42.8**	**23.6**	**43.2**	**51.2**

(a) **Performances under different data representation of bounding box regression targets**: the proposed General distribution supervised by DFL improves favorably over the competitive baselines.

n	Δ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
12	1	40.1	58.4	43.1	23.1	43.8	52.5
14		**40.2**	58.3	**43.6**	**23.3**	44.2	52.2
16		**40.2**	**58.6**	43.4	23.0	**44.3**	**53.0**
18		40.1	58.1	43.1	22.6	43.9	52.6

(b) **Varying n by fixing $\Delta = 1$ on ATSS (w/ GFL)**: The performance is robust to a range of n according to its target distribution in Fig. 5(c).

n	Δ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
16	0.5	**40.2**	58.4	43.0	22.3	43.8	**53.1**
	1	**40.2**	**58.6**	**43.4**	**23.0**	**44.3**	53.0
	2	39.9	58.3	42.9	22.5	43.8	51.8
	4	39.8	58.5	42.8	22.8	43.4	52.3

(c) **Varying Δ by fixing $n = 16$ on ATSS (w/ GFL)**: Small Δ usually leads to better performance whilst $\Delta = 1$ is good enough for practice.

Table 2: Study on DFL (ResNet-50 backbone). All experiments are reproduced in mmdetection [3] and validated on COCO `minival`.

Experiments

QFL	DFL	FPS	AP	AP_{50}	AP_{75}
		19.4	39.2	57.4	42.2
✓		19.4	39.9	58.5	43.0
	✓	19.4	39.5	57.3	42.8
✓	✓	19.4	**40.2**	**58.6**	**43.4**

Table 3: **The effect of QFL and DFL on ATSS**: The effects of QFL and DFL are orthogonal, whilst utilizing both can boost 1% AP over the strong ATSS baseline, without introducing additional overhead practically.