Cascade RetinaNet: Maintaining Consistency for Single-Stage Object Detection

Hongkai Zhang, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen

{hongkai.zhang, hong.chang, bingpeng.ma, shiguang.shan, xilin.chen}@vipl.ict.ac.cn

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

Motivation

Recent single-stage detectors suffer from severe inconsistency problems:

- Classifier is confused by **misaligned classification and localization** due to the unreasonable IoU threshold.
- **Feature inconsistency**: the refined anchors are associated with the feature extracted from the previous location.

Analysis in Inconsistency

Misaligned Classification and Localization:

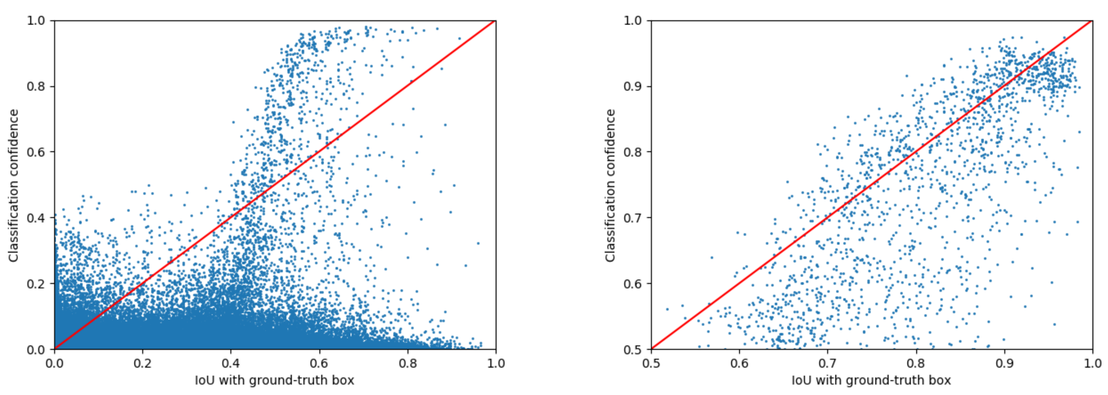

Fig 1: Classification confidence vs. IoU in different cascade stages.

- (left) Classification scores are not well aligned with the IoU for the first stage, especially for the confidences near IoU@0.5.
- (right) Improved consistency between classification and regression in the second stage using increased IoU threshold.

Feature Inconsistency:

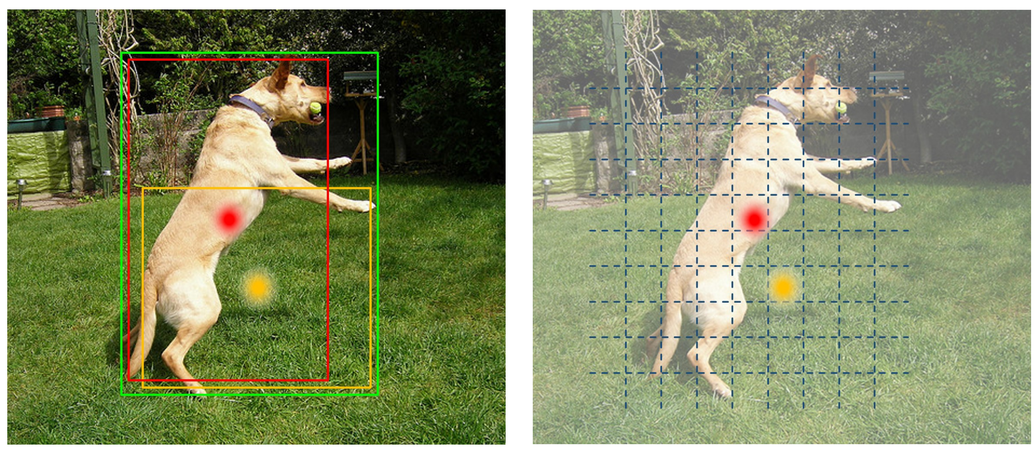

Fig 2: Feature misalignment between original and refined anchor.

- (left) Original image. The green box stands for the ground truth and the orange one represents the original anchor. The refined anchor is shown as the red bounding box.
- (right) Feature grid. Locations of center points for original and refined anchors are plot. It indicates that simply extracting features from the previous location (orange point) is inaccurate.

Proposed Method

Designing rules for the cascade manner:

- Improving consistency between classification confidence and localization performance.
- Maintaining feature consistency between different stages.

Cascade RetinaNet:

- Gradually increase the foreground IoU thresholds to maintain the consistency between classification and localization.
- Encode the current localization information into the features of next stage by Feature Consistency Module (FCM).

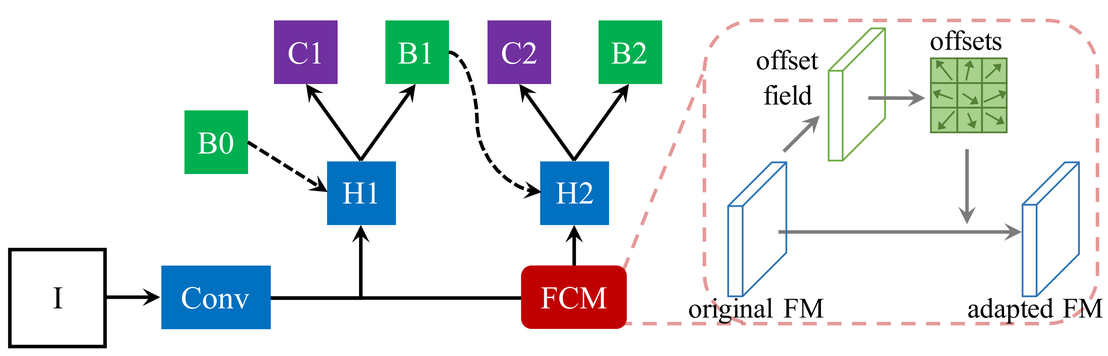

Fig 3: The overall architecture of Cascade RetinaNet.

Experiments

Ablation study:

Method	Scale	IoU	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
RetinaNet [20]	600	-	34.0	52.5	-	-	-	-
Cas-RetinaNet	600	0.5	33.8	52.3	48.1	41.5	29.8	11.2
Cas-RetinaNet	600	0.6	34.4	52.5	48.5	41.9	30.5	11.7
Cas-RetinaNet	600	0.7	34.4	52.0	48.1	41.7	31.3	**12.5**
RetinaNet [20]	800	-	35.4	53.9	-	-	-	-
Cas-RetinaNet	800	0.5	35.4	54.6	50.4	43.0	31.4	11.8
Cas-RetinaNet	800	0.6	**36.1**	**55.0**	**50.8**	**43.9**	**32.5**	**12.5**

Tab 1: Ablation study on IoU thresholds.

- Naively adding a new stage with the same setting brings no gains.
- Increasing the foreground IoU for the second stage is beneficial since it leads to a more consistent distribution.

Backbone	Scale	FCM	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
ResNet-50	600		34.4	52.5	48.5	41.9	30.5	11.7
ResNet-50	600	✓	35.5	54.0	49.7	43.3	32.0	12.6
ResNet-50	800		36.1	55.0	50.8	43.9	32.5	12.5
ResNet-50	800	✓	37.1	56.3	52.2	45.3	33.5	12.8
ResNet-101	800		37.9	56.8	52.8	46.0	34.9	13.9
ResNet-101	800	✓	**38.9**	**58.1**	**53.9**	**47.1**	**36.2**	**14.3**

Tab 2: Ablation study on FCM.

- Steadily improvements under different settings are achieved due to the effectiveness of the adapted feature produced by FCM.

#Stages	Test stage	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
1	1	34.0	52.5	-	-	-	-
2	$\overline{1 \sim 2}$	**35.5**	**54.0**	**49.7**	**43.3**	**32.0**	12.6
3	$\overline{1 \sim 2}$	35.0	53.1	49.1	42.5	**32.0**	12.6
3	$\overline{1 \sim 3}$	34.9	52.9	49.0	42.4	31.9	**12.7**

Tab 3: Ablation study on the number of stages.

- Cascading three stages leads to a slight drop in the overall performance while achieves the best for high IoU. It is a trade-off between sample quality and quantity as mentioned in Cascade R-CNN [1].

Overall performances:

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [20]	ResNet-50	35.7	55.0	38.5	18.9	38.9	46.3
RefineDet512 [35]	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
GA-RetinaNet [3]	ResNet-50	37.1	56.9	40.0	20.1	40.1	48.0
RetinaNet [20]†	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
ConRetinaNet [16]†	ResNet-101	40.1	59.6	43.5	23.4	44.2	53.3
CornerNet511 [17]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	**53.9**
Cas-RetinaNet	ResNet-50	37.4	56.6	40.7	20.9	40.3	47.5
Cas-RetinaNet†	ResNet-101	**41.1**	**60.7**	**45.0**	**23.7**	**44.4**	52.9

Tab 4: Overall performances on COCO `minival` set.

Conclusion

- Analysis shows that **inconsistency in single-stage detectors** is the key factor limiting the detection performance.
- Two main designing rules for **maintaining consistency** are proposed: *improving consistency between classification and localization*, and *maintaining feature consistency between different stages*.
- **Cascade RetinaNet**, a simple but effective architecture, can maintain the consistency by increasing thresholds and adopting FCM, which leads to improved detection performance.