Joint COCO and LVIS workshop at ECCV 2020: LVIS Challenge Track Technical Report: CenterNet2

Xingyi Zhou[1,2], Vladlen Koltun[2], and Philipp Krähenbühl[1,2]

[1]UT Austin, [2]Intel Labs

Abstract. We design a center-based multi-stage detector, CenterNet2. It augments CenterNet with cascade classification and bounding box refinement stages using RoIAlign. The resulting detector combines recent advances in both one-stage and two-stage object detection, while retaining simplicity and speed. CenterNet2 outperforms its CascadeRCNN counterpart by 3 percentage points in bounding box mAP on LVIS. When equipped with a dedicated federated loss and other enhancements, our final model (team CenterNet) achieves 36.1 mask mAP on LVIS validation, surpassing the strong challenge baseline by 9 mAP points.

1 Introduction

Objects detection has seen a shift from anchor-based [6–8] to anchor-free methods [11, 13, 14]. Anchor-free approaches detect objects as points [13] or fixed-size blocks [11] and then regress the object size and a refined location. This simplifies detection and generally provides some speed-up. However, it comes at a cost. All current anchor-free detectors use a one-stage approach and thus miss out on much of the progress in two-stage detection [1, 9].

In this report, we present the first anchor-free multi-stage detector, called CenterNet2. CenterNet2 detects objects by finding their center and regressing their size. For each detection, CenterNet2 extracts region-level features with RoIAlign. These features help refine the detection through a cascade of classifiers [1]. Compared to multi-stage anchor-based detectors, our anchor-free approach is simpler, faster, and more accurate. CenterNet2 does not rely on a region proposal network (RPN), but rather finds objects as point predictions that are later refined. This allows the network to use fewer proposals in RoI heads (256 vs. 1k), leading to faster inference. Compared to single-stage anchor-free methods, our multi-stage detector is more accurate and effortlessly extends to instance segmentation. The later stages make full use of years of progress in multi-stage detection. Furthermore, the carefully-sampled foreground classification head allows applying a more dedicated classification loss. This is critical for long-tailed classification.

CenterNet2 outperforms its RPN-based CascadeRCNN counterparts by 1 mAP on COCO and 3 mAP on LVIS, while being faster. When equipped with a dedicated federated loss, CenterNet2 outperforms the baseline by 4.2 mAP on LVIS. Our final model with a stronger backbone achieves 36.1 mask mAP on LVIS validation and 35.8 mask mAP on LVIS test-dev, surpassing the strong challenge baseline by 9 mAP.

2 CenterNet2: A Multi-stage CenterNet

CenterNet2 builds on an FPN version of CenterNet [13]. A multi-scale keypoint detector first finds object centers in a densely predicted heatmap. For each center, the detector then extracts a local feature at that point and regresses to the size of the bounding box containing the object. CenterNet then simply returns these boxes as potential detections. In CenterNet2, we refine those detections further. For each detection, we extract a region feature using ROIAlign from which we predict a refined classification and bounding box regression. We repeat this procedure for a cascade of region features [1].

While the overall framework is simple, there are subtle details that make a difference. Specifically, CenterNet produces higher quality proposals than a vanilla RPN, thus changes the positive-negative statistics for the RoI head. The conventional Faster/ Cascade RCNN sets a maximal of $\frac{1}{4}$ bounding boxes as positive samples, where the positive is defined by bounding box IoU greater than 0.5 with any ground-truth boxes. Empirically, we observe that a direct adoption of these hyperparameters yields much fewer negative samples and did not improve the performance. To compensate for the change in foreground and background ratio, we increased the positive IoU threshold and NMS threshold for the proposal network to allow more "negative" samples. The resulting positive ratio is set to match the original ratio of $\frac{1}{4}$.

3 Federated Loss for Federated Datasets

LVIS annotates images in a federated way [2], and images are thus only sparsely annotated. This leads to much sparser gradients, especially for rare classes [10]. On one hand, if we treat all unannotated images as negatives, the resulting detector will be too pessimistic and ignore rare classes. On the other hand, if we only use annotated images the resulting classifier will not learn a sufficiently strong background model. Furthermore, neither strategy reflects the natural distribution of positive and negative labels on a potential test set. To remedy this, we choose a middle ground and apply a *federated loss* to a subset S of classes for each training image. S contains all positive annotations, but only a random subset of negatives. We sample the negative categories in proportion to their square-root frequency in the training set, and empirically set $|S| = 50$ in our experiments. During training, we use a binary cross-entropy loss on all classes in S and ignore classes outside of S. The set S is sampled per iteration. The same training image may be in different subsets of classes in consecutive iterations.

4 Experiments

Implementation details. We implement our method based on detectron2 [12]. We base CenterNet2 and an anchor-based baseline on CascadeRCNN [1]. During training, we set the CenterNet loss weights to 0.5 as CenterNet is originally trained with learning rate 0.01. We set the positive IoU threshold to $0.6, 0.7, 0.8$ for each CascadeRCNN stage, and set the CenterNet NMS threshold to 0.9. During testing, we multiply the proposal score by the classification score and change the final NMS threshold to 0.7.

	LVIS v1				COCO	
	AP^{box}	AP^{box}_r	AP^{box}_c	AP^{box}_f	AP^{box}	runtime
CenterNet-FPN [13]	-	-	-	-	39.6	53ms
CascadeRCNN [1]	24.0 $\pm$0.10	7.6 $\pm$0.10	22.9 $\pm$0.15	32.7 $\pm$0.05	42.1	70ms
CenterNet2-specific	26.6 $\pm$0.00	12.8 $\pm$0.30	25.2 $\pm$0.05	34.4 $\pm$0.05	**43.1**	77ms
CenterNet2-agnostic	**26.9** $\pm$0.05	**12.4** $\pm$0.15	**25.4** $\pm$0.15	**35.0** $\pm$0.05	42.9	60ms

Table 1: Results of different detectors on both LVIS v1 validation and COCO validation set (separate models). We use the standard softmax classification for two stage-detectors. All models are ResNet50-1x with FPN P3-P7 and multi-scale training. For LVIS, we report mean and standard deviation over 2 runs. Runtimes are measured using the COCO model on a local Titan X GPU.

We experiment with two versions of CenterNet2: category-agnostic proposals and category-specific proposals. The first version predicts a single-channel heatmap for all categories like an RPN. The category label is then inferred in later stages. The second version predicts bounding boxes from CenterNet in a class-specific manner.

Baseline CascadeRCNN. The default CascadeRCNN uses FPN levels P2-P6. To make it compatible with one-stage detectors [6,13], we change the FPN level to P3-P7. We use the RetinaNet anchor generators: 3 scales $\times$ 3 aspect ratios per level. This modification gives a slight improvement in both runtime and accuracy for bounding box detection on COCO (from 41.9 mAP/ 84ms to 42.1 mAP/ 70ms with ResNet50-1x).

Both the baseline and CenterNet2 use repeat-factor sampling [2] on LVIS. Unless specified, we use the default training size, augmentation, learning rate, and schedule: we use max edge 1333, short edge from 640 to 800, base learning rate 0.02, and train for 90k iterations with learning rate dropped by 10x at the 60k and 80k iterations.

4.1 Results

For simplicity, we conduct ablation studies with bounding box detection. Table 1 compares CenterNet2 to CascaceRCNN and CenterNet baselines on both COCO and LVIS datasets using a fairly small ResNet50 backbone with a $1\times$ schedule. CenterNet2 improves the COCO mAP by 3.5 (category-specific) to 3.3 (category-agnostic) over the baseline CenterNet, and surpasses the RPN-based CascadeRCNN, by 1.1 and 0.8 COCO mAP, respectively. The improvement is even more pronounced on the LVIS dataset with 2.6-2.9 mAP. This highlights the advantage of using a stronger proposal network.

Federated loss. Table 2 compares the proposed federated loss to baselines including the LVIS v0.5 challenge winner, the equalization loss (EQL) [10]. For EQL, we follow the authors' settings in LVIS v0.5 to ignore the 900 tail categories. Switching from the default softmax to sigmoid incurs a slight performance drop. However, our federated loss more than makes up for this drop, and outperforms EQL and other baselines significantly.

Challenge model. Finally, we apply all known tricks to improve CenterNet2-agnostic with federated loss for the LVIS challenge (instance segmentation). The step-by-step results are shown in Table 3. We add a standard Mask RCNN [3] head, train longer (180k

Detector	Loss	AP^{box}	AP^{box}_r	AP^{box}_c	AP^{box}_f
CenterNet2-agn.	Softmax-CE	26.9 $\pm$0.05	12.4 $\pm$0.15	25.4 $\pm$0.15	35.0 $\pm$0.05
	Sigmoid-CE	26.6 $\pm$0.00	12.4 $\pm$0.10	25.1 $\pm$0.05	34.5 $\pm$0.10
	EQL [10]	27.3 $\pm$0.00	15.1 $\pm$0.45	25.9 $\pm$0.20	34.2 $\pm$0.35
	FedLoss (Ours)	**28.2** $\pm$0.05	**18.8** $\pm$0.05	**26.4** $\pm$0.00	**34.4** $\pm$0.00
CascadeRCNN	Softmax-CE	24.0 $\pm$0.10	7.6 $\pm$0.10	22.9 $\pm$0.15	32.7 $\pm$0.05
	Sigmoid-CE	23.3 $\pm$0.10	8.2 $\pm$0.30	21.9 $\pm$0.40	31.5 $\pm$0.25
	EQL [10]	25.7 $\pm$0.01	15.5 $\pm$0.25	24.6 $\pm$0.70	31.5 $\pm$0.45
	FedLoss (Ours)	**27.1** $\pm$0.05	**16.1** $\pm$0.10	**26.0** $\pm$0.35	**33.0** $\pm$0.25

Table 2: Ablation experiments on different classification losses on LVIS v1 validation. We show results with both our proposed detector (top) and the baseline detector (bottom). All models are ResNet50-1x with FPN P3-P7 and multi-scale training. We report mean and standard deviation over 2 runs.

	AP^{mask}	AP^{mask}_r	AP^{mask}_c	AP^{mask}_f	AP^{box}	AP^{box}_r	AP^{box}_c	AP^{box}_f
CenterNet2-agn.	-	-	-	-	28.2	18.8	26.4	34.4
+ Mask	25.3	15.8	24.4	30.6	28.6	17.7	27.0	35.1
+ 2x	27.4	19.0	26.2	32.4	30.6	20.7	28.8	36.9
+ FPN2-6	28.2	20.2	26.7	33.4	31.5	22.6	29.3	37.8
+ X101	30.3	21.2	28.7	36.0	33.9	24.4	31.4	40.8
+ DCN	32.1	22.5	31.1	37.4	35.9	25.2	34.0	42.6
+ PointRend	34.0	25.5	32.2	39.9	36.7	27.2	34.1	43.7
+ larger input	34.9	24.6	33.1	41.4	37.3	24.6	33.1	41.4
+ test-aug	36.1	24.9	34.7	42.5	38.5	26.6	36.6	45.8

Table 3: Technical components of the challenge model. All models are trained from the standard ImageNet initialization. The results are reported based on a single run.

iterations), switch back to FPN level P2-P6 (this slows down the detector significantly), and add a larger backbone network (ResNeXt101). This model yields 30.3 mask mAP, which outperforms the official baseline (27.2 ± 0.17) by 3.1 mAP points. The improvement stems from our contributions: 1) We use CenterNet2 as the detector while the baseline uses a standard Mask RCNN, and 2) we handle imbalanced classification by a federated loss while the baseline uses two-phase training [4].

We then additionally add deformable convolutions [15], use PointRend [5], and use larger image resolution (960×1600) following Autoassign [14]. These enhancements together improve the results to 34.9 mask mAP. Finally, we apply standard [12] test-time augmentation with test scale 400 to 1200 and flip test. The final single model achieves a mask mAP of 36.1 on validation and 35.8 on test-dev. This is the model we submitted to the challenge server.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (2018)
2. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
4. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)
5. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR (2020)
6. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. ICCV (2017)
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
9. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. In: CVPR (2020)
10. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: CVPR (2020)
11. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: ICCV (2019)
12. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
13. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv:1904.07850 (2019)
14. Zhu, B., Wang, J., Jiang, Z., Zong, F., Liu, S., Li, Z., Sun, J.: Autoassign: Differentiable label assignment for dense object detection. arXiv:2007.03496 (2020)
15. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets v2: More deformable, better results. In: CVPR (2019)